1-1-2001

Development and evaluation of test assembly procedures for computerized adaptive testing.

Frederic Robin
University of Massachusetts Amherst

DEVELOPMENT AND EVALUATION OF TEST ASSEMBLY PROCEDURES FOR COMPUTERIZED ADAPTIVE TESTING

A Dissertation Presented

By

FREDERIC ROBIN

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

February 2001

School of Education
Educational Policy Research and Administration
Research and Evaluation Methods Program

DEVELOPMENT AND EVALUATION OF TEST ASSEMBLY PROCEDURES FOR COMPUTERIZED ADAPTIVE TESTING

A Dissertation Presented

By

FREDERIC ROBIN

Approved as to style and content by:

Stephen G. Sireci, Chair

Ronald K. Hambleton, Member

Hariharan Swaminathan, Member

Mzamo P. Mangaliso, Member

Bailey W. Jackson, Dean
School of Education

DEDICATION

To my wife, Tamara,

who sparked and nurtured my inspiration and encouraged me to always do my best.

To my family on both sides of the Atlantic.

ACKNOWLEDGMENTS

I would like to thank my advisor, Stephen G. Sireci, for his guidance throughout my studies at UMASS and also for his constant availability and friendship. As one of his two first doctoral students to graduate, I hope his experience through the process was as good as mine. I am also greatly indebted to Ronald K. Hambleton, Hariharan Swaminathan and Stephen G. Sireci not only for their teaching but also for the excellence of their mentorship. I could always count on them to help me move forward on the many issues and problems that have puzzled me along these four years of studies. I would also like to extend my gratitude to Mzamo P. Mangaliso for his participation in my committee.

I wish to express my appreciation to my fellow students and also to Peg Louraine, Joanne Provost and Betti Swasey with whom important things were discussed and goodies eaten. I am especially grateful to Dehui Xing and Kevin C. Meara for their friendship and support. With them, the computer lab and the Newman center became lively places of scientific and not-so-scientific argument as well as places of wisdom.

I also am very grateful to Babacar Mboup and Giray Berberoglu to whom I owe the opportunity to be part of the Research and Evaluation Methods Program and who have always offered me support and friendship.

ABSTRACT

DEVELOPMENT AND EVALUATION OF TEST ASSEMBLY PROCEDURES FOR COMPUTERIZED ADAPTIVE TESTING

FEBRUARY 2001

FREDERIC ROBIN, INGENIEUR, NATIONAL INSTITUTE OF APPLIED SCIENCES, LYON, FRANCE

M.S., NORTH CAROLINA STATE UNIVERSITY

Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Stephen G. Sireci

Computerized adaptive testing provides a flexible and efficient framework for the assembly and administration of on-demand tests. However, the development of practical test assembly procedures that can ensure desired measurement, content, and security objectives for all individual tests, has proved difficult. To address this challenge, desirable test specifications, such as minimum test information targets, minimum and maximum test content attributes, and item exposure limits, were identified. Five alternative test assembly procedures where then implemented, and extensive computerized adaptive testing simulations were conducted under various test security and item pool size conditions. All five procedures implemented were modeled based on the weighted deviation model and optimized to produce the most acceptable compromise between testing objectives.

As expected, the random (RD) and maximum information (MI) test assembly procedures resulted in the least acceptable tests—producing either the most informative

but least secure and efficient tests or the most efficient and secure but least informative tests—illustrating the need for compromise between competing objectives. The combined maximum information item selection and Sympson-Hetter unconditional exposure control procedure (MI-SH) allowed for more acceptable compromise between testing objectives but demonstrated only moderate levels of test security and efficiency. The more sophisticated combined maximum information and Stocking and Lewis conditional exposure control procedure (MI-SLC) demonstrated both high levels of testing security and efficiency while providing acceptable measurement. Results obtained with the combined maximum information and stochastic conditional exposure control procedure (MI-SC) were similar to those obtained with MI-SLC. However, MI-SC offers the advantage of not requiring extensive preliminary simulations and allows for more flexibility in the removal or replacement of faulty items from operational pools.

The importance of including minimum test information targets in the testing objectives was supported by the relatively large variability of test information observed for all the test assembly procedures used. Failure to take this problem into account when test assembly procedures are operationalized is likely to results in the administration of sub-standard tests to many examinees. Concerning pool management, it was observed that increasing pool size beyond what is needed to satisfy all testing objectives actually reduced testing efficiency.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS v

ABSTRACT vi

LIST OF TABLES x

LIST OF FIGURES xi

CHAPTER

1. INTRODUCTION 1

    1.1 Statement of Problem 3
    1.2 Purpose of the Study 6

2. LITERATURE REVIEW 9

    2.1 Computer-Based Testing 9
    2.2 Measurement 12

        2.2.1 Ability Estimation 13
        2.2.2 Information and Standard Error of Measurement 15

    2.3 Test Content 17
    2.4 Test Security 20

        2.4.1 Item Exposure 22
        2.4.2 Test Overlap 23
        2.4.3 Item Selection and Exposure Control 24

    2.5 Test Efficiency 30

        2.5.1 Pool Management 31
        2.5.2 Test Assembly 32

    2.6 Summary 37

3. METHODOLOGY 38

3.1 Test Assembly Procedures 38
3.2 Testing Situation 41
3.3 Simulations 45
3.4 Evaluation Criteria 46

4. RESULTS 50

4.1 Measurement 50

4.1.1 Test Information 51
4.1.2 Measurement Error and Ability Estimation 58
4.1.3 Summary 61

4.2 Content 63
4.3 Security 63

4.3.1 Preliminary Simulations 63
4.3.2 Item Exposure 65
4.3.3 Test Overlap 67
4.3.4 Summary 73

4.4 Efficiency 74

5. SUMMARY AND CONCLUSIONS 77

5.1 Limitations of the Study 80
5.2 Directions for Further Research 81
5.3 Conclusion 82

BIBLIOGRAPHY 84

LIST OF TABLES

Table		Page
3.1	Measurement Specifications	42
3.2	Content Targets, WDM Specifications, and Pool Content	43
3.3	Item Parameter Statistics	44
4.1	Overall Test Information Statistics	52
4.2	Overall Measurement Error Statistics	59
4.3	Maximum Unconditional and Conditional Item Exposure Rates	66
4.4	Average Peer-to-peer and Test-retest Overlap Exposure Rates	70
4.5	Overall Evaluation of Testing Efficiency	75

LIST OF FIGURES

Figure		Page
3.1	Pool Information and Target Test Information	45
3.2	Comparison of Cumulative Density Functions Between Population and Sample Distributions for Expected and Unexpected Cases	47
4.1	Minimum Test Information and Minimum Test Information Targets	54
4.2	Standard Deviation of Test Information	55
4.3	Proportion of Tests Providing the Minimum Required Information	56
4.4	Mean Test Information	57
4.5	Observed Bias and Standard Error of Measurement	60
4.6	Extreme Test Information Functions for Tests Administered to Examinees at the Same Ability Level (0.00)	62
4.7	Evolution of the Maximum Item Exposure Rates Over Preliminary Simulation Runs	64
4.8	Unconditional Item Exposure Distributions	68
4.9	Conditional Item Exposure Distributions	69
4.10	Peer-to-peer Overlap Distributions	71
4.11	Test-retest Overlap Distributions	72

CHAPTER 1

INTRODUCTION

Traditionally, large-scale high-stakes standardized tests have been designed to be administered in a linear paper-pencil format to large groups of examinees at fixed dates. The development and administration of these tests, generally referred to as linear tests, has been well established. First, test items are created, screened, pre-tested, and screened again for psychometric quality before being incorporated into an operational item bank. Second, a small number of equivalent test forms are assembled from the item bank to satisfy desired measurement targets and test specifications. Third, at a scheduled date, test forms are administered to large groups of examinees and data are collected. Fourth, responses are scored, examinee traits are estimated, and scores are reported. Administered test forms are then retired and the entire development process is repeated for the next scheduled test administration.

Many educational testing organizations and credentialing/licensing agencies have been successful in developing and administering their tests to large numbers of examinees using the linear approach described above. Paper-pencil linear tests demonstrated their value for providing valid, reliable and fair information for the assessment and placement of military recruits, the admission of students to colleges and universities, and the attribution of professional credentials. However, new demands for more flexible and more efficient testing approaches have emerged. In an increasingly fast paced world the availability of tests on-demand, allowing examinees to chose their testing date on any business day, and the reduction of testing time have become

important to most test consumers including students, admission officers, policy makers, professionals, employers, or human resource managers.

Naturally, the traditional linear testing design can be extended to accommodate on-demand testing, so that examinees can schedule their test on any business day. Instead of testing large groups of examinees on few occasions, testing can be spread over time on an individual basis. But then, because of the huge number of items that would be necessary, test forms cannot be retired after each administration. Test forms, or at least a large number of items, have to be reused over time, making it possible for examinees to gain knowledge about the content of the test before its administration. Testing programs that have adapted the linear approach to high-stakes on-demand testing have been able to do so by reassembling new test forms frequently and by always having a number of alternative forms available for random assignment at any time. However, the level of test security that can be obtained with this extension of the linear testing approach to on-demand testing is seriously limited by the amount of resources available for creating new items and for developing new forms. Moreover, testing efficiency is degraded as testing time is not reduced and more resources are required—many more items need to be created and pretested, and more frequent form assembly operations need to be conducted to ensure only limited testing security.

Alternatively, research and development efforts pioneered by Lord (1970, 1977, 1980) and Weiss (1973, 1982) on both item response theory (IRT) and adaptive testing have opened up new test assembly and test delivery approaches that are particularly appropriate for on-demand testing. Taking advantage of computer technology, testing programs such as the Armed Service Vocational Aptitude Battery (ASVAB) for the

examination of military recruits to assign them to training school or job specialties (Sands, Waters, & McBride, 1997), the Graduate Record Examination (GRE) for the admission of candidates to graduate school (Eignor, Way, Stocking, & Steffen, 1993), and the National Council Licensure Examination (NCLEX-RN and NCLEX-PN) for the licensing of registered and practical/vocational nurses (Zara, 1994) have re-engineered their operations and began offering computerized adaptive testing (CAT) versions of their tests in 1997, 1993, and 1992, respectively. As a result of their success in producing and administering shorter on-demand tests comparable in quality to their predecessors, these programs have created a tremendous interest for CAT as a new approach to test design and delivery. Credentialing programs such as those offered by Novell (Foster, Olsen, Ford & Sireci, 1997) to millions of professionals around the world quickly followed and many existing testing programs such as those offered by Microsoft, the Law School Admission Council, or the American Institute for Public Accountants are now either already conducting research to assess the feasibility of CAT for their own purpose or close to switching to CAT or other computer-based testing (CBT) designs appropriate for on-demand testing.

1.1 Statement of Problem

Although progress has been made in developing, administering, and maintaining testing programs operating on-demand, important challenges remain to ensure the desired level of quality for all examinees (National Council on Measurement in Education Ad hoc Committee on Computerized Adaptive Test Item Disclosure, 1996; Drasgow, 1998; Hambleton, in press). In particular, appropriate choices of test design

and test assembly, administration, and scoring methodologies that can support clearly defined testing objectives must be made. Simulation studies to evaluate and optimize fully automated computerized adaptive testing systems must be conducted. A monitoring of the testing objectives must be put in place. And, frequent maintenance operations must be prepared and executed.

Expanding on Davey and Parshall's description of the on-demand testing challenge (Davey & Parshall, 1995), four types of testing objectives can be identified for the development, evaluation, and monitoring of on-demand testing programs: (1) measurement precision, (2) content balancing, (3) security, and (4) efficiency of test administrations. Unfortunately, these objectives are at odds with each other. For example, achieving a higher level of test security generally results in lower measurement precision and/or the need for more items and items of higher quality. Therefore, the formulation and prioritizing of these objectives and the choice of test assembly and test delivery methods are crucial. A particular testing program may not reach its desired level of quality if its objectives are improperly specified and/or may not be sustainable over time if its design and procedures require too many items, too frequent maintenance operations or do not have sufficient flexibility to produce acceptable compromise between competing objectives.

A review of the literature pertinent to on-demand testing reveals that much of the attention has been focused on developing methodologies for improving specific aspects of ability estimation, test assembly and/or control of item exposure. Less attention has been focused on defining comprehensive criteria for evaluating the respective quality of testing designs and applications, and on studying interactions

between measurement, content, security and efficiency objectives. Only a few studies (Stocking & Lewis, 1995,1998; Chang & Twu, 1998; or Chang, Ansley, & Lin, 2000; for example) have reported comprehensive results obtained under high-stakes conditions where test specifications were comprehensive (i.e., including content constraints and conditional exposure constraints). As a result, testing programs who want to start testing afresh or want to switch from linear to on-demand testing are faced with a very large choice of designs and procedures which are largely untested under their specific conditions.

Another challenge that arises for test developers is that most of the procedures (e.g., ability estimation, item selection, and item exposure control) used in test assembly, administration, and scoring require optimum specifications and settings to be determined. Unlike traditional linear tests, for which each possible form is known and has been evaluated well in advance of administration, individualized CBT forms are assembled in real-time. Consequently, the composition of an examinee test will depend on many factors such as the examinee responses, the population of examinees to be tested or the examinees who have been tested before, the pool of items available for selection, as well as on the test specifications, and on the test assembly procedures and their settings. Therefore, preliminary simulation studies are necessary to determine the feasibility of the testing objectives for the target population and the resources available, and to establish the most appropriate test specifications and test assembly settings (including minimum and maximum test length, test content attributes and weights, examinee prior ability distribution, and item exposure parameters, for example). Inappropriate settings or procedures may result in much poorer performance than

expected on one or more of the test objectives and may jeopardize the overall quality and efficiency of the testing programs. Unfortunately, if large organizations have developed their own software, to the author's knowledge, no comprehensive simulation software has been made available to the public.

Also, the issue of measurement consistency across examinee tests appears to have been largely overlooked in published reports. In most studies, estimates of test reliability and average conditional test information values are reported, but the variability of test information and the minimum level of test information provided to examinees are rarely reported. Again, unlike with the traditional linear test approach, for which measurement properties such as test information are determined and validated well in advance for all examinees, CBT test forms which are assembled in real-time cannot be expected to have the same properties across examinees, and not even across those at the same ability level. Consequently, with no mechanism to ensure a minimum level of test information, it is possible that some examinees are not provided with an adequate opportunity to demonstrate their ability, despite acceptable average test information levels for the target population (Davey & Fan, 2000). This problem is likely to be prevalent for highly constrained tests (i.e., tests that include complex and restrictive content and exposure specifications) assembled from item pools of limited sizes and/or quality.

1.2 Purpose of the Study

Although the problems discussed in the previous section are largely related to one another, they were addressed separately. First, a review of the literature was

conducted to identify the most important designs and procedures for conducting CBT and to select appropriate criteria for evaluating their quality with regard to each of the four testing objectives previously mentioned.

Second, a computer program was developed to simulate the administration and scoring of computer-based tests (CBTs) in a wide range of testing situations. This computer program should help test developers evaluate the demands of conducting on-demand testing. It should also help test developers to choose among a number of alternative designs and procedures and determine the size and quality of item pools and the algorithmic settings (such as item exposure limits and content attribute penalty weights) that will ensure the realization of their testing objectives in their particular testing situation.

Third, a simulation study was conducted to evaluate strengths and limitations of current test assembly procedures in ensuring consistent realization of measurement objectives across examinee tests. More specifically, the effect of item exposure control on overall test information and the variability of test information across examinees were investigated under a variety of test assembly methods and testing conditions. Thus, important factors affecting the level of measurement precision obtained for each examinee such as item pool size, item selection procedure (item selection plus exposure control), and item exposure limits were manipulated. Also, because exposure control procedures may not perform as expected when the target examinee population used in setting up exposure control parameters differs from the actual examinee population, additional simulations were conducted under unmatched target and actual examinee populations.

Fourth, following up on the issue of measurement consistency, the potential of alternative item selection strategies and optimization methods for ensuring a minimum level of test information for all examinees were discussed. Further directions for research were provided.

CHAPTER 2

LITERATURE REVIEW

In this chapter, the basic steps and the main approaches for conducting computer-based testing are presented and specific methodologies and procedures for computerized adaptive testing test assembly are reviewed.

2.1 Computer-Based Testing

Computer technology is revolutionizing testing in the same way it has been revolutionizing the media. Computers offer a new medium for creating, storing and delivering items, collecting and scoring examinee responses and scoring examinee tests. Beyond the traditional multiple-choice paper-pencil items which can be easily transferred onto the computer, new item formats that make use of on-line documentations, audio and video materials, and simulated interactions have been made available, expanding the range behaviors, aptitudes and skills that can be assessed (Parshall, Davey & Pashley, in press; Zenisky, 2000). Computers also offer what they have initially been built for, the power to automate tasks and make expert decisions without human intervention, making it possible to adapt and deliver tests on-demand to large numbers of examinees.

Assuming that enough items have been created, validated, and made available to the computer, designs and methodologies for assembling, delivering, and scoring high quality tests to examinees are needed. Such designs and methodologies are the focus of this literature review.

Given a relatively large set of items, generally referred to as an item pool, made available to the system for selection, the basic steps for assembling, delivering and scoring computer-based tests can be described as follows (Mills, & Stocking, 1996; Sands et al., 1997; Wainer, Dorans, Flaugher, Green, Mislevy, Steinberg, & Thissen, 1990):

Step 1. Initialize the system for a new examinee to be tested

Step 2. *Select an item (or a set of items, or a test form)*

Step 3. Administer the item

Step 4. Capture the examinee response and score the item

Step 5. Update the examinee ability estimates and the test characteristics

Step 6. *Decide to continue or terminate the examinee test*

a) Continue testing by going back to Step 2

b) Terminate testing by going to Step 7

Step 7. Finalize the examinee proficiency estimate and report test results

Step 8. Start testing a new examinee by going back to Step 1

Most of these steps simply describe the automation of the test delivery, administration, and scoring tasks. However, Steps 2 and 6 (italicized) may involve expert decisions built into what is generally referred to as the computer-based testing (CBT) system in the form of heuristics and/or optimization algorithms. Several families of approaches or designs have been proposed for building CBTs.

With computerized linear testing (CLT), an approach equivalent to traditional paper-pencil linear testing, parallel forms are assembled in advance. Each test

administration is then determined by the selection (generally in a random fashion) of one of the available forms (Step 2). In this case, the computer is used simply as a sophisticated test administration medium that automates test form assignment, data collection and scoring (Steps 1, 3 ,7). The main advantage of CLT over traditional testing is the possibility of using new item formats.

However, taking full advantage of computer power, Lord (1977) demonstrated that it was possible to use the information collected on the examinee as testing progresses to adapt the test to the examinee and as a result significantly improve the efficiency of the testing process. By using item response theory (IRT) (Hambleton & Swaminathan, 1985; Lord, 1980; Lord & Novick, 1968), he had overcome the major problem of producing comparable scaled scores (i.e., scores defined on a reference scale) despite administering different sets of items and thus different tests to different examinees. Following Lord's computerized adaptive testing (CAT) approach, no pre-assembled forms are necessary. For each examinee, items are selected one at a time (Step 2) from the entire set of items (item pool) made available for testing until the test is completed. At each stage of administration, the item selection strategy employed evaluates the amount of IRT information that may be provided by any of the items available and selects the most informative item for administration.

Other approaches to computer-based testing have been developed that fit between the two extremes exemplified by CLT and CAT. Computerized linear on-the-fly testing (COFT) is equivalent to CAT without the adaptive part. COFT is appropriate for generating randomly parallel forms using classical test theory (Gibson & Weiner,

1998) or IRT, and may be a viable alternative when item scoring cannot be automated (e.g., when essay items are used).

Computerized multi-stage testing (CMST) is an intermediate approach between CLT and CAT in the sense that it uses pre-assembled modules (forms smaller than the test length) instead of items to adapt the test to the examinee. For each examinee, modules are selected one at a time (Step 2) from the set of modules available at any particular testing stage until the test is completed (Lord, 1980; Luecht & Nungester, 1998). The CMST design allows for more control over the content of each test and the individual test assembly is simplified (since much of the work is done earlier when the modules are assembled) (Patsula, 1999). However, because there is fewer decision points at which the test can be adapted to the examinee, its potential for the optimization of test assembly is less than that of CAT and the extent to which test security can be controlled is more limited.

In the remainder of this chapter, attention is focused on test assembly methodologies for CAT. Alternative approaches and procedures to assemble computerized adaptive tests and ensuring the realization of measurement, content, security, and efficiency objectives are reviewed.

2.2 Measurement

Ability, item information and standard error of measurement estimates play an important role in test assembly and examinee scoring. At any testing stage, the best possible estimate of the examinee's ability and the amount by which any available item may contribute to the reduction of measurement error are needed before the next item

selection decision can be made. The last estimates obtained at the end of each test administration can then be used for reporting examinee scores. Ability estimates and test information values can also be used for evaluating the quality of tests at both individual and group levels.

2.2.1 Ability Estimation

Because it is simple to implement and because it has desirable asymptotic properties (Hambleton & Swaminathan, 1985; Lord & Novick, 1968), the maximum likelihood estimation (MLE) with known item parameters is the most commonly used procedure for estimating examinee's ability. Assuming both unidimensionality of the underlying construct measured and local independence of item responses (Hambleton, Swaminathan, & Rogers, 1991), tests can be modeled using the well known three parameter logistic model (3PL) which expresses the probability of correct or incorrect responses to an item i as a function of examinee's latent ability θ and item characteristics a_i (discrimination), b_i (difficulty), and c_i (guessing) by

$$P(X_i \mid \theta; a_i, b_i, c_i) = \left\{ \begin{array}{l} c_i + (1-c_i)\dfrac{e^{1.7a_i(\theta-b_i)}}{1+e^{1.7a_i(\theta-b_i)}}, \text{if } X_i = 1 \text{ (correct response)} \\ 1-c_i-(1-c_i)\dfrac{e^{1.7a_i(\theta-b_i)}}{1+e^{1.7a_i(\theta-b_i)}}, \text{if } X_i = 0 \text{ (incorrect response)} \end{array} \right\}. \quad (2.1)$$

Ability estimates (or score), $\hat{\theta}$, for any examinee who has responded to $i = 1,\ldots,n$ items can then be found by computing the mode of the likelihood function

$$L(\theta \mid X_i, a_i, b_i, c_i; i = 1, n) = \prod_{i=1}^{n} P(X_i \mid \theta). \quad (2.2)$$

Alternatively, Bayesian approaches have been developed that may resolve some problems encountered with maximum likelihood estimation (MLE), namely the impossibility of finding estimates for perfect (all correct or all incorrect examinee responses) or near perfect responses and the relatively large standard error of measurement obtained when only a few responses are available (Owen, 1975; van der Linden, 1998a). Given prior knowledge on the examinee distribution expressed by the function $f(\theta)$, $\hat{\theta}$ can also be computed using expected a-posteriori (EAP) estimation. In this case, $\hat{\theta}$ represents the expectation of θ over the posterior ability distribution obtained from combining the likelihood and the prior functions as follows:

$$\hat{\theta} = E[P(\theta \mid X_i, a_i, b_i, c_i; i = 1, n)] \tag{2.3}$$

where

$$P(\theta \mid X_i, a_i, b_i, c_i; i = 1, n) = \frac{\prod_{i=1}^{n} P(X_i \mid \theta) \cdot f(\theta)}{\int_{-\infty}^{\infty} \left(\prod_{i=1}^{n} P(X_i \mid \theta) \right) \cdot f(\theta) \cdot d\theta} . \tag{2.4}$$

Although, when it was developed, EAP estimation was too computer intensive to be used without making approximations (Owen, 1975), the full EAP procedure described here has now become an attractive alternative for estimating and updating examinees' ability during test administration (Step 5) and possibly for deciding when to stop testing (Step 6) when tests are allowed to vary in length. However, less biased and often equally accurate towards the end of the test, the MLE may still be preferred for computing final examinee scores at the end of each test administration (Wang & Vispoel, 1998).

2.2.2 Information and Standard Error of Measurement

Naturally, ability estimates have a degree of precision that can be estimated in the form of standard error of measurement. Two types of procedures may be used to evaluate the standard error of measurement associated with any examinee test and to evaluate the potential contribution of any new item for reducing measurement error.

Using the asymptotic properties of MLE, the standard error of measurement associated with $\hat{\theta}$ can be estimated as

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\hat{\theta})}} \tag{2.5}$$

where $I(\hat{\theta})$ represents the amount of information provided by the test that can be used for evaluating the examinee's ability (Hambleton et al., 1991). Given that item parameters have been established, test information can be easily obtained from item information. At any ability point, test information is

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) \tag{2.6}$$

where

$$I_i(\theta) = \frac{2.89 a_i^2 (1 - c_i)}{\left[c_i + e^{1.7 a_i (\hat{\theta} - b_i)}\right]\left[1 + e^{-1.7 a_i (\hat{\theta} - b_i)}\right]^2} \tag{2.7}$$

represents the contribution of item *i* to the total test information.

From Equations 2.6 and 2.7 it can be seen that test information is a function of the items administered which is independent of the examinee responses (van der Linden, 1998a). Therefore, the standard error provided by Equation 2.5 can only reflect

the level of measurement precision that can be expected from a test independently from any examinee to whom it might be administered to. Thus, estimates of the standard error of measurement obtained based on test information are referred to as test standard error of measurement or *TSEM* .

Alternatively, using a Bayesian approach, estimates of the standard errors of measurement may be obtained based on the variance of the posterior ability distribution. Referred to as examinee standard error of measurement, or *ESEM*, these estimates can be computed by

$$ESEM(\theta) = \sqrt{Var\left[P(\theta \mid X_i, a_i, b_i, c_i; i = 1, n)\right]}. \qquad (2.8)$$

This approach does take into account both the item characteristics (in the form of calibrated item parameters) and the examinee responses pattern. Thus, *ESEMs* may be more appropriate for estimating individual standard errors of measurement, as examinees responding with unexpected response patterns cannot be measured as well as those who do fit the model reasonably well.

Measurement error estimates play an important role in score reporting because they provide necessary information for the interpretation of test scores. In most CAT algorithms, they are also essential for selecting the most appropriate items for administration. IRT information, in particular, has been found to be very convenient for item selection because of its additive nature and because of its computational simplicity. Items providing the most information at the current ability estimate, among all the items available for selection, are likely to increase test information the most and are consequently the most desirable candidates for selection and administration. This strategy for item selection, referred to as the maximum information strategy (MI), was

proposed by Lord (1977, 1980). The work done by researchers such as Owen (1975) or van der Linden (1998a) on Bayesian approaches, and Chang and Ying (1996) on global information approaches to item selection have shown that promising alternatives are available, but more work remains to be done to fully evaluate their utility in practical cases, in particular when content and security considerations must also play an important part in the item selection.

In terms of evaluating the quality of the tests produced (quality control), the distinction between test information and examinee information is important. Because test information is independent of examinee responses, it can be understood as the opportunity given to examinees to demonstrate their ability. In CAT, not only are examinees of different ability administered different tests, but examinees of the same ability are also administered different tests. Consequently, the opportunity provided to examinees to demonstrate their ability will vary. It is important that items and tests be fair to all examinees. Therefore, it is important that examinees be provided similar or at least a minimum level of opportunity to demonstrate their ability, regardless of how well they take advantage of that opportunity. Thus, although it is desirable to use the most accurate ability estimates (*EAP* and *ESEM*) for test assembly, IRT information is better suited for evaluating and ensuring test quality and for deciding when to stop testing in variable test length situations.

2.3 Test Content

Content specifications are used in the item selection process to ensure each test includes the required number of items and is assembled with the desired balance of

categorical and quantitative attributes (Stocking & Swanson, 1993; Swanson & Stocking, 1993; van der Linden & Boekkooi-Timminga, 1989; van der Linden & Reese, 1998). Test attributes are simply the sum of the attributes associated with the items included in the test. Content specifications may be fairly simple, including for example only test length (fixed or variable) and basic content blueprints, or more complex, including multilevel content blueprints, items sets and item enemy requirements, testing time limits, and speededness conditions as well. Generally, categorical item attributes are set as discrete parameters indicating the item's status: 1 if the item possesses the attribute or 0 otherwise. Quantitative item attributes, on the other hand, are set as scalar parameters reflecting descriptive or statistical item properties such as response time mean and standard deviation (van der Linden, Scrams & Schnipke, 1999), discrimination, word count, or cost.

Relationships between test specifications and item attributes can be expressed in terms of constraints that must be satisfied for any test to be content valid. Defining integer decision variables x_i for each item i most content specifications can be written as

$$\sum_{i=1}^{N} x_i a_i = s,$$

$$\sum_{i=1}^{N} x_i a_i \leq \text{or} \geq s, \text{ or}$$

$$s_L \leq \sum_{i=1}^{N} x_i a_i \leq s_U, \tag{2.9}$$

where $x_i = 1$ if item i is included in the test or $x_i = 0$ otherwise; a_i represents the value of an item attribute; and s or s_L and s_U define an attribute specification (exact value to be met, or lower and upper bounds). Similarly, item sets (or testlet) decision variables can also be used to guide the selection of groups of items (all or part of the items related to a common stimulus, for example).

Thus, procedures can be developed to ensure that tests exactly meet all specifications, or, in cases where it is not possible, come as close as possible to them as in the weighted deviation model (WDM) proposed by Stocking and Swanson (1993). With the WDM, a weighting of the different categorical and quantitative specifications may be used to reflect their relative importance and to influence the item selection towards the smallest possible sum of weighted deviations. Any specification constraint (2.9) for which deviations from the ideal targets are to a certain degree acceptable can be rewritten as

$$s_L \leq \sum_{i=1}^{N} x_i a_i + d_L - d_U \leq s_U , \tag{2.10}$$

where the slack variables (smallest positive reals) d_L and d_U represent deviations from the lower and upper attribute specifications, respectively. Then, for any item selection, the sum of weighted deviations over $k = 1,...,K$ weighted deviation type specifications can be expressed as

$$\sum_{k=1}^{K} w_{Lk} d_{Lk} + \sum_{k=1}^{K} w_{Uk} d_{Uk} \tag{2.11}$$

where w_{Lk} and w_{Uk} represent the under and over realization penalty weights associated with attribute k and where $d_{Lk} \geq 0$ and $d_{Uk} \geq 0$ are determined from K (2.10) equations.

Hence, whenever content specifications cannot be satisfied or when limited but acceptable departure from ideal test content attributes result in better overall tests, the items violating the least important content specifications (lower weights) will be preferred over those that would violate more important ones (higher weights).

Evaluation and monitoring of the agreement with content specifications can be done simply from an account of the proportion of tests that do not satisfy all specifications and from the average weighted deviations overall and by content specification. Through repeated simulations during the development phase of the test assembly procedure, successive adjustments can be made to the weights in order to minimize deviations from the content specifications and improve the performance of the procedure in the specific testing situation at hand.

2.4 Test Security

The validity of a testing program may be compromised when some examinees are able to gain confidential knowledge about the test they are about to take (Davey & Nering, 1998; Sands, Waters & McBride, 1997; Zara, 1997). Clearly, examinees who have practiced with or have been coached on items that will be on the test have an unfair advantage over those who have not. Less obvious but also troublesome are situations where specific item features (particularly with story based or simulation

items) become known or when inappropriate strategies to "beat the test" are spread around or taught.

Typically, pre-knowledge about tests may be obtained in two important ways: (a) through security breaches, or (b) through normal exposure of the items. Security breaches such as unintended disclosure, or theft of items or test forms, can be prevented by putting in place tight security controls on test development, test delivery and test administration processes. In addition, testing programs usually keep alternative test forms ready to become operational in case a security breach is discovered.

Item exposure becomes a problem when items are reused over time, as is the case with on-demand testing. Although examinees are prevented from taking any test materials when they leave the test, they still may remember the content of the test for a period of time after the examination. Examinees may then reconstruct with some degree of accuracy items that will be part of future tests. This information could be communicated to future examinees. As was demonstrated by the Kaplan/ETS incident, the higher the stakes, the higher the temptation and the ingenuity spent by examinees and coaching organizations to obtain pre-knowledge about the tests.

Important progress has recently been made in improving test security through the control of item exposure. A number of indicators for evaluating security risks related to item exposure have been proposed and new methodologies for controlling exposure have been developed, particularly in the context of computerized adaptive testing (CAT) (Davey & Nering, 1998; Robin, 1999a; Stocking & Lewis, 1998).

Clearly, the more frequently items are administered the more likely it is their security will be compromised. Thus, straightforward indicators such as item exposure

(i.e., the ratio of the number of times an item has been administered as part of a test by the total number of tests administered) and average item exposure are very useful in evaluating security risk. However, these indicators are insufficient to detect and control exposure patterns that have been found to significantly increase security risks. Additional indicators, such as conditional item exposure, test overlap, test-retest overlap and pair-wise exposure, should also be monitored. A general description of most of these exposure indicators is provided below.

2.4.1 Item Exposure

Using x_{ai} to keep track of the exposure ($x_{ai} = 1$) or non-exposure ($x_{ai} = 0$) of item i to examinee a, and given that N examinees have been tested, overall item exposure can be computed by

$$X_i = \frac{\sum_a x_{ai}}{N}. \qquad (2.12)$$

Average item exposure does not require the knowledge of each item exposure value. It can be obtained simply based on the number of items in the pool, n, the number of items administered to each examinee, k_a, and the total number of examinees tested, N; that is

$$\bar{X} = \frac{\sum_a \frac{k_a}{N}}{n} \quad \text{or} \quad \bar{X} = \frac{k}{n} \text{ if test length is fixed to } k. \qquad (2.13)$$

With the Kaplan/GRE incident it was discovered that maintaining overall item exposure to desirably low levels is not enough for maintaining security (Stocking & Lewis, 1995, 1998). Because adaptive testing seeks to match examinee ability with item difficulty (Lord, 1980) items tend to be used exclusively with examinees whose

ability level is close to the item difficulty. As a result, item exposure will be much higher amongst examinees of similar ability (conditional item exposure) than it appears when all examinees are taken into account. With N_h representing the number of examinees tested in conditional ability group h ($h = 1,...,G$), the conditional item exposure values for any item i can be computed in the same way item exposure values were computed, that is

$$X_{ih} = \frac{\sum_{a \in h} x_{ai}}{N_h}, \quad h = 1,...,G \tag{2.14}$$

2.4.2 Test Overlap

Another angle from which to evaluate test security is test overlap (Davey & Parshall, 1995). Test overlap (or between-test overlap, to be more precise) can be defined as the proportion of items in common between any two tests. The greater the test overlap, the greater the security risk, because examinees are able to share more information. Average test overlap estimates can be easily obtained from item exposure when test length is fixed. For all examinees as well as for any given examinee in group h, Chen, Ankenmann & Spray (1999) showed that average conditional overlap (O_h) can be estimated by

$$O_h = \frac{\sum_{i=1}^{n} X_{ih}^2}{k}, \text{ or} \tag{2.15}$$

$$O_h = \frac{1}{\bar{X}_h} S_X^2 + \bar{X}_h, \tag{2.16}$$

where $\overline{X}_h$ and S_X^2 are the average and variance of item exposure. However, because distributions with very different shapes (e.g., oppositely skewed) may have similar average values, average test overlap may not be sufficient to thoroughly evaluate security risks. Estimates of the test overlap distribution may be obtained by sampling from the set of all observable overlap values obtained from any two administered tests.

Peer-to-peer overlap represents the overlap between any two examinee tests (unconditional overlap), while test-retest overlap represents the overlap between any two tests administered to the same examinee assuming the examinee's ability has not changed (conditional test overlap). Test-retest overlap (Davey & Parshall, 1995) is particularly important when examinees are allowed to retake the test within a short period of time. In this case, the likelihood that examinees will remember significant portions of the test and be able to focus their preparation for the next administration is greatly increased. Test-retest overlap can be evaluated by keeping track of test overlap between examinees at the same ability level (conditional test overlap).

2.4.3 Item Selection and Exposure Control

Generally, item selection and item exposure control procedures work together by having: (1) the item selection routine rank all the items available for selection according to their desirability (amount of information at the theta estimate, for example); (2) the exposure control routine reject the possible selection of some of the items according to some exposure criteria; and (3) the best of items still available selected for administration. Ideally, item exposure control procedures should only

reject items (especially the ones that have been ranked as the most desirable) that would become overexposed if selected.

Early exposure control procedures were designed primarily to spread exposure from highly desirable to less desirable items. In the "4-3-2-1" randomesque procedure (McBride & Martin, 1983), for example, the first item to be administered is selected randomly from the best 4 items available (accepted with a probability of .25 and rejected with a probability of .75), the second one from the 3 best available, the third one from the 2 best, the fourth and all subsequent ones are then simply the best available (acceptance and rejection probabilities of 1.0 and 0.0, respectively). This procedure is very simple to implement but its effectiveness has been shown to be very limited and no real control of exposure is provided (Chang & Twu, 1998; Revuelta & Ponsoda, 1998).

The Sympson and Hetter (SH) procedure (Hetter & Sympson, 1997; Sympson & Hetter, 1985) improved on the randomesque procedure by using an appropriate probabilistic exposure control parameter for each item. In practice, given $K_i,\ i = 1,\ldots,n$, item exposure parameters, the SH procedure can be applied as follows. For any examinee, even before the test administration starts (Step 1 of CAT item selection, initialization):

1. Generate n uniform (0,1) random numbers, $r_i, i = 1,\ldots,n$, and
2. Make unavailable for administration any item for which $K_i \geq r_i$.

Then, the normal item selection procedure can be applied and the procedure is repeated for the next examinee. The SH item exposure control parameters are determined though a series of iterative simulations where examinees' abilities are sampled from the

expected population distribution. Initially, all K_i are set to 1 (probability of acceptance of 1.0). Then, at the end of each simulation run, observed item exposures are compared with the exposure limit set in advance and K_i, $i = 1,...,n$, are adjusted to lower values for items with observed exposure above the limit and to higher values for items with observed exposure below the limit. Simulations are repeated until all K_i incrementally converge to stable values between 0 and 1.

The SH procedure provides effective control over item exposure rates, but fails to provide control over more complex patterns of exposure such as conditional item exposure or test overlap. It should also be noted that the SH item exposure parameters have to be re-estimated each time the item pool is modified, and that the procedure may not be as effective as expected if the distribution of simulated examinees used to establish the appropriate item exposure control parameters does not match the real examinee population.

The Kaplan/GRE incident, in which a concerted effort by a group of high ability examinees resulted in a very serious security breach, demonstrated the insufficiency of unconditional procedures to ensure security and sparked new developments in exposure control. In particular, Stocking and Lewis (1995, 1998), and Davey and Parshall (1995) proposed new procedures that provide additional control over conditional item exposure rates and that, at least theoretically, are not sensitive to mismatch between expected and observed examinee population distributions. Also based on probabilistic item exposure control parameters, these conditional procedures require extensive preliminary simulations to determine their appropriate values. With n items available in the pool and h ability-conditional examinee groups, the Stocking and Lewis conditional (SLC)

procedure requires the determination of *nxh* parameters to ensure that observed conditional item exposures remain below the specified limits.

In practice, the SLC procedure can be applied as follows. For any examinee at any stage of test administration:

1. Obtain an ordered list of the *m* most desirable items available for selection from the item selection strategy.
2. Form the operand probability of administration, k_i^*, for each item *i* in the list by computing

$$k_i^* = \left\{\prod_{j=1}^{i-1}(1-K_{jl})\right\}K_{il}, \qquad (2.17)$$

where K_{il} is the exposure control parameter associated with item *i* given that the current examinee ability estimate belongs to ability level *l*.

3. Adjust the operand probabilities, by dividing them by their unadjusted sum, so that they sum to one. Form the corresponding cumulative distribution.
4. Generate a uniform (0,1) random number.
5. Select the item that corresponds to the random number in the cumulative distribution and make all the preceding ones unavailable for further selection with the current examinee.

Note that with SLC, exposure cannot be controlled at the test initialization step (Step 1 of CAT item selection) because the ability level of the examinee tested may change as new items are administered, thus changing ordered list in step 1 above.

The preliminary simulations needed for establishing the appropriate SLC item exposure parameters are conducted in a similar fashion as that of the SH procedure.

However, although conditional exposure control results in the reduction of test overlap, SLC does not provide a mechanism for imposing limits on test overlap rates. Preliminary simulations may be repeated until the most appropriate maximum exposure limit specifications yield satisfactory item exposure and test overlap rates.

The Davey and Parshall (DP) procedure (Davey & Nering, 1998; Davey & Parshall, 1995) differs from the SLC procedure by additional conditioning on the items previously administered within test administration. The DP procedure requires the determination of at least n item exposure and $n(n-1)/2$ pair-wise item exposure parameters to ensure that item exposure and overlap are below acceptable limits.

Alternatively, Revuelta and Ponsoda (1998) and Robin (1999a) proposed stochastic unconditional and conditional exposure control procedures. With these procedures, items are prevented from being selected based on comparisons between observed exposures, computed from past administrations, and specified exposure limits. After each test administration, observed item exposure rates (unconditional and/or conditional) are updated, and items with exposure rates above the specified limits are prevented from further administration. These items are made available again as soon as their observed exposures come back within limits (after new examinees have been tested) while other items that have become overexposed are made unavailable.

Given S_l, $l = 1,...,L$, specified item exposure limit, the stochastic conditional (SC) procedure proposed by Robin (1999a) can be applied as follows. For any examinee at any stage of test administration:

1. Obtain the conditional ability level l to which the current estimate belongs (more than one level could be obtained when overlapping exposure control intervals are used).
2. Make unavailable for administration any item for which $K_{il} \geq S_l$.

At the end of each test administration, the stochastic item exposure control parameters are updated as

$$K_{il}^{*} = K_{il} + \frac{x_i - K_{il}}{T}, \quad i = 1,\ldots,N \tag{2.18}$$

where $x_i = 1$ if item i was administered and 0 otherwise, and where T is set to control the rate at which stochastic parameters are allowed to change.

The SC procedure has the practical advantage of not requiring extensive simulations in order to set the item exposure control parameters. These can be initially set to values above exposure limits. Then, with T set to a reasonably small value (say 30), the stochastic parameters are quickly updated to effective values after relatively few test administrations. Thus, pools can be more easily maintained (updated or changed).

Unfortunately, although the case for conditional exposure control has been clearly made, very few studies have compared the respective merits of conditional procedures under realistic conditions including both comprehensive test specifications and exposure limits. The simulation study conducted by Chang and Twu (1998) is probably the most comprehensive study available. In this study, comparisons of the randomesque, SH, unconditional Stocking & Lewis (SL), SLC and DP procedures were made with a CAT under medium size and large item pools (360 and 720 items for 30 item tests) and few content specifications (6 content areas). Overall, SLC and DP were

successful in ensuring desirable levels of security. Both succeeded in maintaining overall exposure rates below .2 and .1, conditional exposure rates at about .4 and .2, and test-retest mean overlap at about .2 and .15, with the medium and large item pool sizes employed, respectively. However, security was obtained at the cost of relatively large and moderate increases in average standard error of measurement with the medium size and large pools, respectively. The other unconditional procedures (randomesque, SH and SLU) were by far unable to provide the desired level of exposure control. Another, study (Robin, 1999a) showed the potential of the stochastic conditional approach to provide effective control over conditional item exposures under highly constrained, high stakes conditions. Because no or very limited preliminary simulations are needed, this approach has the appeal of simplicity and practicality. But more research is needed to evaluate its effectiveness comparatively with other procedures.

2.5 Test Efficiency

In practice, the creation, pretesting, and screening of items is a very expensive and time-consuming process. As a result, the number and the quality of items that can be made available for assembling tests are limited, which in turns limits the quality of the tests that can be administered. Therefore, the challenge is to produce tests of the desired quality and at the same time make the best use of the available items. Consequently, test efficiency can be viewed and defined as the capacity to produce needed tests, i.e., tests that consistently (for all examinees) satisfy all measurement, content, and security objectives, while requiring the smallest number of items to be

made available. Hence, test efficiency can be operationalized as the ratio of the number of items produced to the number of tests administered.

Leaving aside item creation, pretesting, and screening, two major operations contribute to test efficiency: pool management and test assembly. A brief review of pool management and a more detailed review of test assembly methodologies is provided below.

2.5.1 Pool Management

Tests are assembled by selecting items from an item pool. One could imagine that all the items available at any point in time could constitute the pool from which tests are assembled. However, to avoid the risk of losing all the items at once, to further improve item security, and to make a better use of all the items, it is more practical and efficient to create parallel item pools from the complete collection of items available which, using ETS' terminology is referred to as the "item vat" (Patsula & Steffen, 1997; Way, Anderson & Steffen, 1998), and to rotate them.

Methodologies for assembling item pools and deciding on their maintenance and rotation schedules have been developed (Guo, Way, & Reshetar, 2000; Stocking & Swanson,1998; van der Linden, Veldkamp & Reese, 1999; Way, & Steffen, 1998; Way, Steffen & Anderson, 1998). In practice, regular pool maintenance operations are scheduled in which operational pools are replaced by new ones. During a maintenance operation, items from the operational pools are either returned to the vat and made available for the assembly of new pools to become operational, set aside for a period of time before being allowed back into the vat, or retired definitely. The rules used for

making those decisions, referred to as "docking rules" at ETS (Guo, Way, & Reshetar, 2000), include: (a) an exposure rate threshold above which items will be set aside and returned to the vat for the next assembly cycle, and (b) a total cumulative item use threshold above which items are retired from the vat. Moreover, the data collected during an administration cycle can be used to conduct item quality control so that flawed items (misskeyed, DIF, drift, etc.) are removed and either revised or deleted.

Clearly, the task of the pool management procedures is to provide adequate pools for the test assembly procedures to produce tests of the desired quality. However, test assembly procedures themselves need to be as efficient as possible for the whole system to be efficient.

2.5.2 Test Assembly

As indicated earlier, tests must satisfy all of the four testing objectives previously outlined according to their operational definitions and specifications. However, as the complexity and the number of test specifications and the stakes for examinees increase, satisfying all measurements, content, and security objectives for all examinees in an efficient way with a limited item pool becomes a very difficult problem.

The typical approach is simply to select items one at a time in a purely sequential fashion. At each testing stage (selection and administration of one item), each item from the pool is evaluated against all the specified measurement, content, and exposure control constraints and either accepted or rejected as a possible candidate for administration. Items should be rejected if they: have already been administered in one

of the previous stages; are overexposed (according to the item exposure control procedure); or contain content attributes that would result in an unbalanced test. Then, the best of the remaining available items is selected for administration. Items are thus selected and administered one at a time until the required number of items has been administered (fixed test length) or until the desired level of measurement precision (variable test length) has been reached and no further testing is needed.

However, this straightforward sequential approach ("greedy heuristic", in the mathematical programming lingo) may fail to provide optimum solutions or even fail to provide any solution for some examinees (Stocking & Swanson, 1993; van der Linden & Reese, 1998). These difficulties generally do not happen when large item pools are available and when few constraints are imposed on the item selection, but they cannot be ignored in most practical situations where complex constraints are needed and item availability is limited (due to exposure control, in particular). Part of the problem comes from the fact that ability estimates are very inaccurate early on and changing as test administration progresses. Although gains in accuracy are dependent on information, the systematic pursuit of highly informative items narrowly focused on the often wrong ability level early in the test is likely to be unproductive when item use is limited (exposure control). Simulation studies conducted by Revuela and Ponsoda (1998), Chang and Ying (1999), and Robin (1999a) demonstrated this phenomenon. Part of the problem also comes from the fact that under complex constraint structures, where items have multiple attributes, some combinations of item selections may exhaust the pool before the normal end of the test administration. For example, items might cover geometry or algebra, but also contain tables or graphics and for some of them

provide clues to others. Consequently, a series of early "unlucky" selections could create a situation in which no more items can be found to satisfy the need for algebra items that do not also contain graphics (already fully covered) or for which clues have not been already given (Stocking & Swanson, 1993; van der Linden & Reese, 1998).

To address these difficulties, test assembly problems can be formulated using a mathematical programming framework similar to that used in automated test assembly (Swanson & Stocking, 1993; Theunissen, 1985; van der Linden & Boekkooi-Timminga, 1989). Mathematical programming models have two major advantages. First, the selection of the next item to be administered can be optimized considering not only the selection of the next item (local optimization) but also the items that remain to be selected until the test is completed (global optimization). Second, more complex optimization functions measuring the desirability of each possible solution (the next items to be administered) can be used that take into account the multiple objectives defining the quality of the test (Robin, 1999a; Veldkamp, 1999).

The weighted deviation model (WDM), proposed by Stocking and Swanson (1993) for severely constrained CAT, is probably the most comprehensive mathematical programming algorithm used operationally in large scale high-stakes testing situations. Using the notations introduced earlier, it can be formulated as follows:

Maximize

$$\sum_{i=1}^{N} x_i I_i(\hat{\theta}) - \sum_{k=1}^{K} w_{Lk} d_{Lk} - \sum_{k=1}^{K} w_{Uk} d_{Uk} \qquad \text{(optimization function)} \quad (2.19)$$

$$s_{Lk} \le \sum_{i=1}^{N} x_i a_{ki} + d_{Lk} - d_{Uk} \le s_U, \quad k = 1,..K \qquad \text{(deviations)} \quad (2.20)$$

Subject to

$$s_{Lm} \leq \sum_{i=1}^{N} x_i a_{mi} \leq s_{Um}, \quad m = 1,\ldots,M \qquad \text{(content constraints)} \quad (2.21)$$

$$x_i = 0 \text{ or } 1, \quad i = 1,\ldots,N \qquad \text{(item exposure control)} \quad (2.22)$$

$$d_{Lk}, d_{Uk} \geq 0 \qquad \text{(WDM variables restrictions)} \quad (2.23)$$

and

$$x_i \in \{0, 1\}, \quad i = 1,\ldots N \qquad \text{(integer decision variables)} \quad (2.24)$$

Equations 2.21 to 2.24 define the set of all admissible tests while equations 2.19 and 2.20 define the relative value of each admissible test. However, despite the fact that a solution to this formulation is a complete test, the sequential nature of the CAT test assembly process remains because the estimate of ability $(\hat{\theta})$ on which any solution depends changes (and improves) as the test administration progresses. To make the best use of all the information available and adapt to examinee responses, a new optimum test has to be found after each item administration. Thus, tests are assembled and administered by repeatedly

1) Finding the best possible "shadow test" including the items already administered and the next items to administer given the current ability estimate
2) Selecting one of the items not already administered for administration
3) Administering the item selected and updating $\hat{\theta}$

(van der Linden & Reese, 1998).

The task of finding the best possible test after each item administration by simple enumeration and ranking of all possible combinations is impossible in practice—with a pool of size n and k items still to be administered the combinatorial complexity of the problem is $n!/[(n-k)!k!]$. However, the heuristic algorithm used by Stocking and Swanson (1993) for solving the WDM has proved to be quite efficient for providing secure high quality tests in highly constrained situations (Stocking & Lewis, 1998). Although not theoretically optimum, the heuristics employed do provide globally optimized solutions for assembling tests in the sequential manner outlined above.

Alternatively, the use of more sophisticated mixed integer programming algorithms for assembling optimum "shadow tests" has been proposed (Armstrong, Jones & Cordova, 1997; Cordova, 1997; van der Linden & Reese, 1998). However, although solutions obtained at each item selection stage are theoretically optimum when the algorithm is allowed to converge, applications have proved computationally challenging in situations where relaxing the constraints was not necessary to ensure feasibility and where no conditional exposure control where enforced to ensure test security.

Offering the means to incorporate complex content specifications, mathematical models and algorithms have tremendously improved CAT designs and applications. However, as attested by the increasingly large number of CAT related sessions presented in recent years at the National Council on Measurement in Education conferences, many problems remain to be better addressed. Variations on the formulation of the test assembly problem and new heuristics for solving them have been proposed that seek to improve the compromise that can be reached between

measurement, content, security, and efficiency objectives. For example, Chang and Ying (1999), and Chang and van der Linden (2000) proposed to add item difficulty pool stratification constraints (a-stratified CAT) as a way to incorporate measurement error into the mathematical programming formulation and as a way to relieve heavy handed item exposure control constraints. With the same goals in mind and with the additional concern for providing equally informative test to equally able examinees, Robin (1999a) and Davey and Fan (2000) proposed to incorporate gradual information targets in the optimization function.

2.6 Summary

Clearly, tremendous efforts have been made to address the numerous problems that have emerged as on-demand CAT and other CBT applications have been implemented. Many alternative designs and procedures have been proposed for solving specific measurement, content, and security problems and for optimizing overall testing efficiency. However, the sheer number of alternatives for conducting CAT and the lack of comprehensive information on their respective merits can be overwhelming.

In this chapter some of the most prominent CAT methodologies and the most important criteria for their comprehensive evaluation were identified. On this basis, a study focusing on the development and thorough evaluation of five alternative test assembly procedures was conducted. The methodology employed and the results obtained are detailed in the following chapters.

CHAPTER 3

METHODOLOGY

A computer program was developed to simulate administration and scoring of computer-based tests in a wide range of testing situations (Robin, 1999b). The purpose of this computer-based testing simulation program (CBTS) is to help test developers evaluate and select the most appropriate design and procedures for their testing situation, optimize their specifications and algorithmic settings, and develop and maintain their operational testing programs.

Using CBTS, a simulation study was conducted to evaluate strengths and limitations of test assembly procedures in ensuring the realization of measurement specifications for all examinee tests under different levels of test security. Comparisons between test assembly procedures were made with respect to multiple testing objectives, where the most highly rated procedures should satisfy measurement and content specifications, ensure the best level of testing security, and make the most efficient use of the whole bank of items available.

The test assembly procedures investigated, the testing situation, simulations, and the criteria for evaluating results are detailed in the following sections.

3.1 Test Assembly Procedures

Five test assembly procedures were investigated. First, random (RD) and maximum information (MI) procedures were used to establish baseline performance. With the random procedure, test items are selected based on uniform random draws

from the subset of items that satisfies all the content constraints. In this way, all the items within each content category have the same probability of being administered, which minimizes both maximum item exposure (conditional and unconditional) and test overlap (peer-to-peer and test-retest) at all ability levels. As a result, RD is expected to perform best with regards to test security. However, RD is expected to perform poorly with respect to measurement precision because no provisions are made for the selection of informative items. With MI, on the other hand, test items are selected almost entirely based on the amount of information they can provide at the examinee's ability estimate, again provided that all the content constraints remain satisfied. As a result, MI is expected to perform well with respect to measurement precision while performing poorly with respect to test security because the number of most informative (discriminating) items over all ability levels is in all practical cases very small compared to the total number of items in the pool.

The three other test assembly procedures investigated included combinations of the maximum information item selection strategy with different item exposure control procedures. A better compromise between measurement and test security was expected by using test assembly procedures that would both seek to select items that can provide high levels of information and avoid items that tend to become overexposed otherwise (multiple-objective optimization). As a result of that compromise, it was also expected that by giving more chances to a larger number of items to be included in test administrations a better use of the item pool would result and higher efficiency would be achieved.

The combined maximum information selection and Sympson-Hetter exposure control (MI-SH) is one of the earlier and most commonly used procedure (Davey & Nering, 1998; Eignor et al., 1993). Experience has shown that this procedure can be effective in providing high measurement precision while sufficiently reducing overall item exposure and making more effective use of item pools in low to moderately high-stakes situations. The combined maximum information and Stocking-Lewis conditional exposure control (MI-SLC) procedure, probably the most powerful test assembly method proposed for highly constrained, high-stakes testing (Stocking & Lewis, 1998), has been shown to provide an effective compromise between high test security and measurement precision demands under such conditions (Chang & Twu, 1998; Stocking & Lewis, 1998). Finally, the combined maximum information and stochastic conditional exposure control procedures (MI-SC) using an improved version of the stochastic conditional item exposure procedure proposed by Robin (1999a) was also included as a potentially simpler and more flexible alternative to the MI-SLC procedure.

Ability estimation and examinee scoring, an essential part of any CAT system, was handled using expected a posteriori (EAP) estimation. To avoid estimation bias, a relatively weak normal prior was used, with mean and standard deviation parameters set to the mean and twice the standard deviation of the target examinee population.

Following common practice in the context of IRT modeling, the ability scale was set for the examinee population to be distributed with a mean of 0.0 and a standard deviation of 1.0.

Initial ability estimates used in selecting the first test item were set by uniform random draws from an ability interval ranging from one standard deviation below the target population mean to the target population mean. This was done to start each test with moderately easy items and at the same time avoid focusing early item selections to a very limited number of items.

3.2 Testing Situation

The testing situation investigated was chosen to exemplify a realistic high-stakes case where item writers can produce only a limited number of items with varying degrees of discrimination, difficulty, and guessing (multiple-choice items) belonging to relatively few exclusive content areas. Realistic testing objectives were specified in the form of measurement and content specifications (including test length, minimum test information targets, and content balancing targets) that each individual test should meet. No security and efficiency specifications were set a-priori. Instead, the best possible compromise between all testing objectives was sought, given satisfaction of measurement and content specifications with the available item pool.

Item pools of different sizes were generated. Two options were investigated, either a number of small size parallel pools or a smaller number of moderately large parallel pools assembled from the whole bank of available items. All things equal, it was to be seen if smaller pools and more frequent rotations would lead to better test efficiency that larger pools and less frequent rotations—the rotation schedule being determined by security considerations.

Tables 3.1 and 3.2 indicate the measurement and content specification values chosen to reflect typical assessments, where reasonable measurement should be provided in a wide range of ability levels and where a balance between content areas should be maintained. Measurement specifications were also set to reflect the minimum amount of IRT information that each test should provide at the final ability estimate. Table 3.1 specifies the measurement targets at each of seven ability levels in term of minimum test information, $TI(\hat{\theta})$, and maximum test standard error of measurement, $TSEM(\hat{\theta})$.

Table 3.1

Measurement Specifications

Ability Estimate	Minimum Test Information Target	Maximum Test Standard Error of Measurement
(-4, -1.61]	5.0	.45
(-1.61, -1.12]	6.0	.41
(-1.12, -0.84]	7.0	.38
(-0.84, 0.84]	8.0	.32
(0.84, 1.12]	7.0	.38
(1.12, 1.61]	6.0	.41
[1.61, 4)	5.0	.45

Table 3.2 indicates the content specifications and the item availability of the two pools for each of six content areas. Content specifications were set in terms of proportion of test items within each content category and in terms of minimum and maximum number of items per content category with a test length fixed at 30 items. The definition of minimum and maximum targets and penalty weights associated with each content attribute, in the manner of the WDM (Stocking & Swanson, 1993), allowed flexibility in the test assembly and prevented its possible failure in assembling tests when items could not be found to exactly match specifications. Equal penalty

weights—1.0 for under-representation and 2.0 for over-representation—were set for each content category, reflecting the fact that no category was considered more important than any other category and that it was considered less acceptable to have not enough than too many items in any content category. The chosen content specifications were in line with practice (Chang & Twu, 1998; Eignor et al., 1993; Stocking & Swanson, 1993). Given sufficiently large and high quality pools, it was expected that desirable compromises between all testing objectives could be obtained.

The two pools were generated to represent the availability of either small or moderately large resources. Pool 1 and pool 2 included 200 and 400 items, respectively. Overall, their content matched the test specifications reasonably well.

Table 3.2

Content Targets, WDM Specifications, and Pool Content

Content Area	Target [a]	WDM Specifications				Pool 1 [d] (200 items)	Pool 2 [d] (400 items)
		w_L [b]	Lower [c]	w_U [b]	Upper [c]		
1	.10	1.0	3	2.0	4	.10	.10
2	.27	1.0	7	2.0	8	.22	.22
3	.17	1.0	5	2.0	6	.20	.20
4	.17	1.0	4	2.0	5	.13	.13
5	.13	1.0	4	2.0	5	.16	.16
6	.17	1.0	5	2.0	6	.19	.19
Total	1.00		28		34	1.00	1.00

[a] In proportion of test items; [b] Penalty weights; [c] Lower and upper targets in number of items for 30-items fixed length tests; [d] In proportion of pool items.

To facilitate comparisons between the use of small and moderately large item pools, pool 2 item parameters and content attributes were generated first and pool 1 was selected as a subset of pool 2. Items in pool 1 were then randomly selected by content

area to maintain the same proportion of items within each area. Item attributes were set randomly with a probability of being assigned to any one of 6 mutually exclusive content areas equal to the corresponding test content proportion specified. True item parameters, a (slope), b (difficulty), and c (guessing), were randomly drawn from LN(0.8,0.25), N(0.0, 1.5), and U(0.0, 0.35) distributions to simulate multiple-choice items with various degrees of discrimination, difficulty, and guessing. Table 3.3 summarizes the item parameter for the two pools.

Table 3.3

Item Parameter Statistics

	Mean	SD	Min	Max
Pool 1 (200 items)				
A	.83	.28	.40	1.68
B	.13	1.22	-2.30	2.73
C	0.17	0.10	0.00	0.34
Pool 2 (400 items)				
A	0.82	0.26	0.40	1.75
B	0.12	1.24	-2.64	2.73
C	0.17	0.10	0.00	0.35

Although it would be more realistic to simulate pretest data and calibrate item parameters and then use the true parameters for simulating item responses and the calibrated parameters for estimating examinees' scores, true item parameters were used for both simulating responses and estimating examinee scores. This approach was preferred here because the purpose of this investigation was to evaluate and compare test assembly methodologies independently from the effect of calibration error.

Figure 3.1 shows the potential of test information offered by each item pool. It also illustrates the minimum test information targets, and by comparison, shows the adequacy of the pools with the measurement objectives.

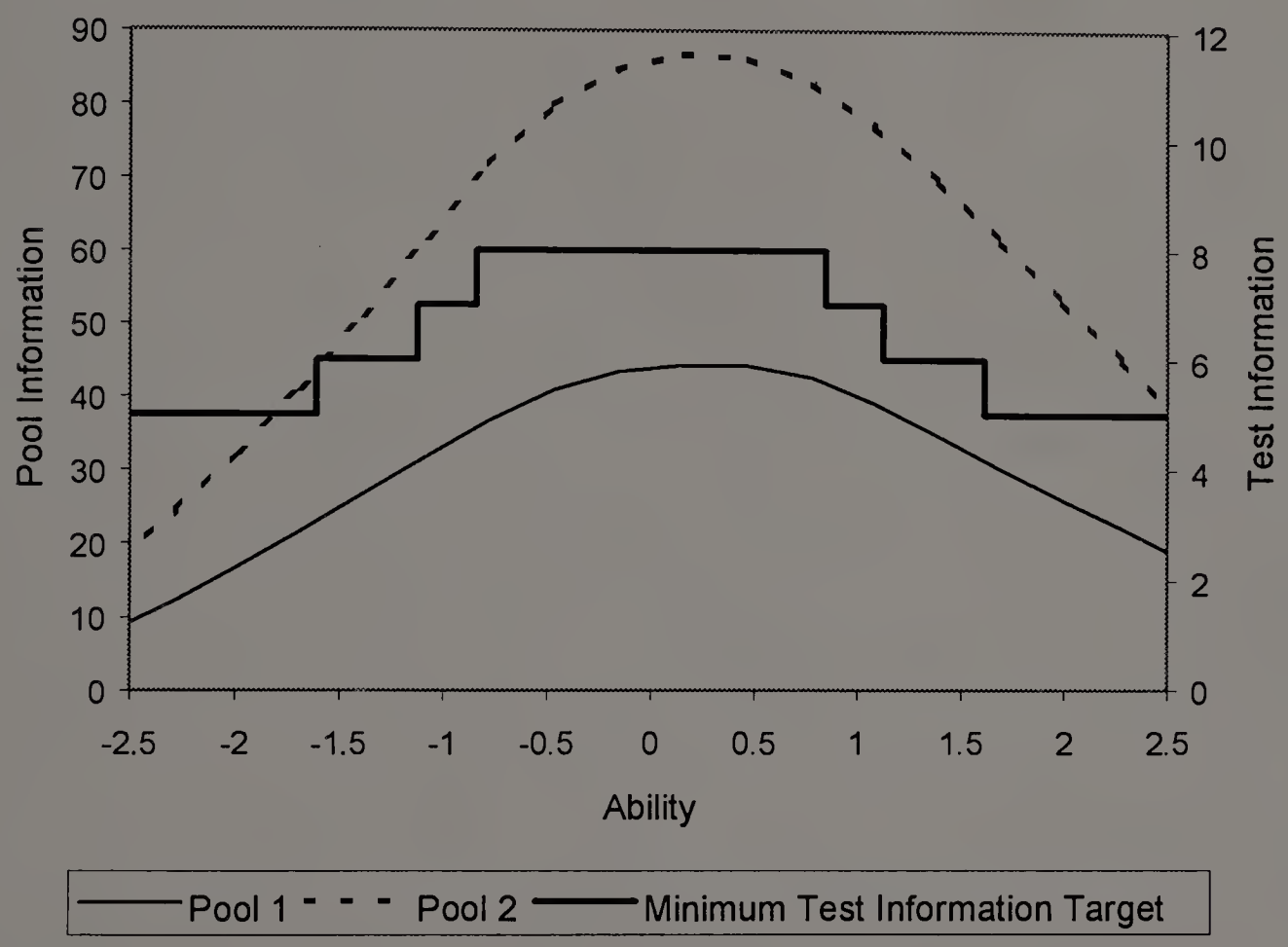


Figure 3.1. Pool Information and Target Test Information. Note: to facilitate comparisons, pool and test information are displayed on different scales.

3.3 Simulations

Four important factors affecting the realization of each testing objective as well as the overall testing performance (i.e., the quality of compromise between testing objectives) were manipulated: pool size (200 and 400 items), match between target population and actual examinee sample distributions (match, and no match with a 0.5 mean difference), test assembly procedure (RD, MI, MI-SH,MI-SLC, MI-SC), and item exposure control level (none, .35, .25 and .18 maximum item exposure). Not all item

exposure control conditions were applicable to all the test assembly procedures, therefore a total of 38 simulation conditions were executed.

In the matched case, simulations were performed by administering the test to large examinee samples including 400 examinees at each one of 15 ability values for replication (6,000 examinee in total). These unequally spaced values, –1.93, -1.28, -0.96, -0.72, -0.52, -0.33, -0.16, 0.0, 0.16, 0.33, 0.52, 0.72, 0.96, 1.28 and 1.93, were computed for the examinee samples to approximate the N(0,1) target population distribution. To facilitate comparisons, the same 15 ability values were used in the unmatched case, but with 106, 210, 225, 248, 277, 304, 317, 338, 384, 424, 453, 522, 593, 793 and 806 examinees to approximate the unmatched actual distribution N(0.5,1). Figure 3.2 shows the degree of approximation between examinee samples and population distributions. Better approximation could have been obtained by increasing the number of ability levels, but 15 were judged to be sufficient. The examinee replications at each ability level were useful for obtaining estimates of conditional results.

3.4 Evaluation Criteria

Each testing simulation was evaluated according to measurement, content, security, and efficiency criteria.

Evaluation of measurement results was based on overall and conditional results at the examinee group level and at the individual level. Estimated standard errors of measurement were obtained through EAP estimation (*ESEM*, equation 2.8) and test information (*TSEM*, equation 2.5). True standard error of measurement (*SEM*) and

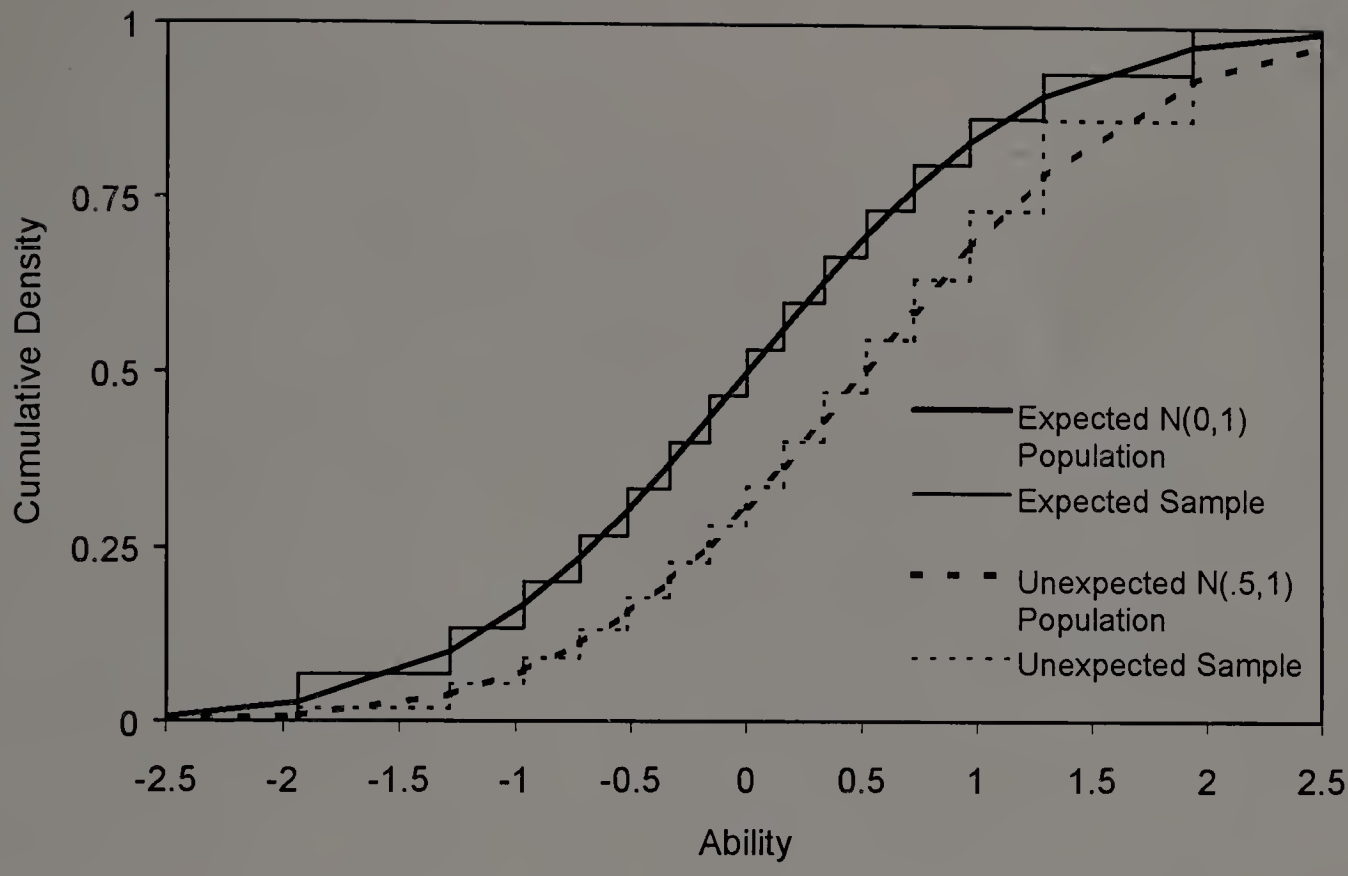


Figure 3.2. Comparison of Cumulative Density Functions Between Population and Sample Distributions for Expected and Unexpected Cases

bias statistics were computed based on true estimation error $(\theta - \hat{\theta})$. Overall statistics were obtained from whole examinee samples. Conditional statistics were obtained from replications at the given θ ability levels for true conditional bias (*CB*) and true conditional standard error of measurement (*CSEM*), and at given $\hat{\theta}$ ability levels for estimated conditional standard error of measurement (*ECSEM* and *TCSEM* using EAP estimation and test information, respectively). Given a=1,…,R replications, true conditional bias and conditional standard error of measurement were computed using

$$CB(\theta) = \frac{1}{R}\sum_{a=1}^{R}(\theta - \hat{\theta}_a) \tag{3.1}$$

and

$$CSME(\theta) = \sqrt{\frac{1}{R}\sum_{a=1}^{R}(\hat{\theta}_a - \bar{\hat{\theta}})^2}\,, \tag{3.2}$$

while estimated conditional standard error were computed using

$$ECSME(\hat{\theta}) = \sqrt{\frac{1}{R}\sum_{a=1}^{R} Var[P(\theta \mid X_a)]} \tag{3.3}$$

and

$$TCSME(\hat{\theta}) = \sqrt{\frac{1}{R}\sum_{a=1}^{R} \frac{1}{I(\hat{\theta}_a)}} \tag{3.4}$$

The true error statistics (not available in real testing situations since true abilities are not known) were used to evaluate the extent to which the estimation procedures used overpredicted true measurement precision under each simulation condition. The extent to which examinees were all provided an opportunity to demonstrate their ability was evaluated based on minimum test information, standard deviation of test information, and proportion of examinees whose test information was above the minimum test information target values.

Evaluation of content was based on the extent to which content specifications (Table 3.2) were satisfied. Average deviations from each content specification were reported. Ideally, all tests should satisfy all content specifications but small deviations could be acceptable.

Evaluation of security was based on the observed exposures and overlap statistics obtained after each simulation run. Maximum item exposures, distribution of item exposure rates, average peer-to-peer and average test-retest overlap rates, and distributions of peer-to-peer and test-retest overlap rates, were reported.

Finally, considering only the procedures and conditions that could produce tests satisfying all measurement, content, and security specifications, evaluations of testing efficiency were made based on the notion that, for a given test length, the ideal pool

usage is obtained when all the items are used with equal frequency (Chang & Ying, 1999). Thus, following these authors' suggestion, testing efficiency was measured by the degree of closeness between observed and ideal item use and computed as

$$\chi^2 = \sum_{i=1}^{N} \frac{(X_i - \overline{X})^2}{\overline{X}}, \tag{3.5}$$

where, X_i and $\overline{X} = \frac{n}{N}$ represent the item and average item exposure rates, respectively. Values close to zero would then indicate high level of testing efficiency and pool usage (as expected with a purely random item selection, for example) and values up to $N(1-\overline{X})$ a poor level of testing efficiency and pool usage (as expected with an item selection strategy aimed solely at maximizing test information, for example).

This index also allows for comparison between alternative pool designs. Given p parallel pools used simultaneously, overall item and average item exposure rates become $X_i^* = \frac{X_i}{p}$ and $\overline{X}^* = \frac{\overline{X}}{p}$ which leads to the overall efficiency index being equal to the pool usage index, i.e., $\chi^{*2} = p\sum_{i=1}^{N} \frac{(X_i^* - \overline{X}^*)^2}{\overline{X}^*} = \chi^2$. As a result, the efficiency of using *2p* pools of size *N* versus *p* pools of size *2N* (the same number of items used overall) can be directly compared by looking at their respective pool usage indexes.

In the end efficient test designs should result in both average test length and χ^2 pool usage index to be low.

CHAPTER 4

RESULTS

The results of this study are organized according to the test objectives previously outlined. Measurement, content, and security results obtained under each one of the 38 simulation conditions outlined in the previous chapter are presented. Then, taking all testing objectives into account, the efficiency of each test assembly procedure is evaluated.

Both overall and conditional group level results are provided. Overall results were compiled over whole examinee samples while conditional results were compiled over a number of ability groups. The overall results provided summary information about the performance of each test assembly procedure that was easier to interpret and more generalizable to other testing situations. The conditional results provided more thorough information useful for the evaluation of the test assembly procedures in the context of testing situation at hand.

4.1 Measurement

The measurement results were reported in terms of test information and measurement error. Mean and standard deviation of test information provided descriptive information useful for evaluating the respective properties of each test assembly procedure. Minimum test information and percentage of tests satisfying the minimum test information targets determined the capacity of pools and test assembly procedures to satisfy the measurement specifications.

Because examinee true abilities were known, the observed measurement errors could be reported. Bias and standard error of measurement results were used to validate the ability, test information and standard error of measurement estimations.

4.1.1 Test Information

Table 4.1 provides the overall test information statistics. As expected RD and MI test assembly procedures lead to the most extreme results. Across simulation conditions, RD consistently (SD of about 1.6) produced poorly informative tests (mean below 6.0) that did not meet the minimum test information targets (only 8% did). MI on the other hand, was less consistent (SD of 2.4 to 3.0) but produced the most informative tests (mean above 15.5) that almost always meet the minimum test information targets (in 99% of the cases or more).

Among the three maximum information and exposure control test assembly procedures under investigations, MI-SH produced the most informative tests that almost always meet the minimum targets. Under the most severe conditions (small pool and .25 unconditional exposure control) test information remained above 10.3 in average, its variability was moderate with a SD of 1.4, and satisfied the minimum targets in 98% of the cases.

Overall, MI-SC and MI-SLC, which both included conditional exposure control, produced similar results. With the small pool under the least severe exposure control condition (conditional exposure limits set to .35) 90% of the tests produced satisfied the minimum targets (relatively low mean and high SD of test information of about 9 and 2, respectively). With the larger pool, both procedures were able to produce satisfactory

Table 4.1

Overall Test Information Statistics (N=6,000)

Test Assembly Method (Exposure Specifications)	Mean	SD	Minimum	% of Tests Above Target
Small Pool (n=200), Expected Sample				
RD	5.5	1.70	0.21	8
MI-SH(.35)	12.2	1.57	1.09	99
MI-SC(.35)	9.2	2.12	0.39	90
MI-SLC(.35)	8.8	2.09	0.33	90
MI-SH(.25)	10.3	1.42	1.44	98
MI-SC(.25)	7.7	2.03	0.42	60
MI-SLC(.25)	6.5	1.84	0.22	29
MI	15.6	2.60	2.37	99
Larger Pool (n=400), Expected Sample				
RD	5.4	1.61	0.16	6
MI-SH(.35)	16.2	1.88	5.17	100
MI-SC(.35)	12.7	2.26	2.16	99
MI-SLC(.35)	12.2	2.20	2.01	99
MI-SH(.25)	14.5	1.44	4.45	100
MI-SC(.25)	11.1	2.15	0.91	98
MI-SLC(.25)	10.0	2.10	0.56	96
MI-SH(.18)	12.4	1.30	4.58	99
MI-SC(.18)	9.0	2.10	0.35	86
MI-SLC(.18)	7.9	1.96	0.52	73
MI	19.2	2.99	4.86	100
Small Pool (n=200), Unexpected Sample (+.5 in mean ability)				
MI-SH(.35)	12.4	1.40	3.41	99
MI-SC(.35)	9.3	1.85	0.68	94
MI-SLC(.35)	9.0	1.89	0.55	93
MI-SH(.25)	10.5	1.22	2.18	99
MI-SC(.25)	7.7	1.79	0.28	66
MI-SLC(.25)	6.6	1.69	0.26	35
Larger Pool (n=400), Unexpected Sample (+.5 in mean ability)				
MI-SH(.35)	16.3	1.69	7.39	100
MI-SC(.35)	12.8	2.02	3.11	99
MI-SLC(.35)	12.3	1.97	3.43	99
MI-SH(.25)	14.6	1.31	6.84	100
MI-SC(.25)	11.2	1.90	1.56	99
MI-SLC(.25)	10.1	1.82	0.87	98
MI-SH(.18)	12.5	1.21	6.02	99
MI-SC(.18)	9.1	1.88	0.53	89
MI-SLC(.18)	8.0	1.69	0.26	79

tests for 99% of the examinees under moderate exposure control (.35), 96% under more severe exposure control (.25), and less than 86% under the most severe exposure control (.18). Mean test information were above 10 points under exposure limits set to .35 and .25. The variability of test information remained large at about 2.

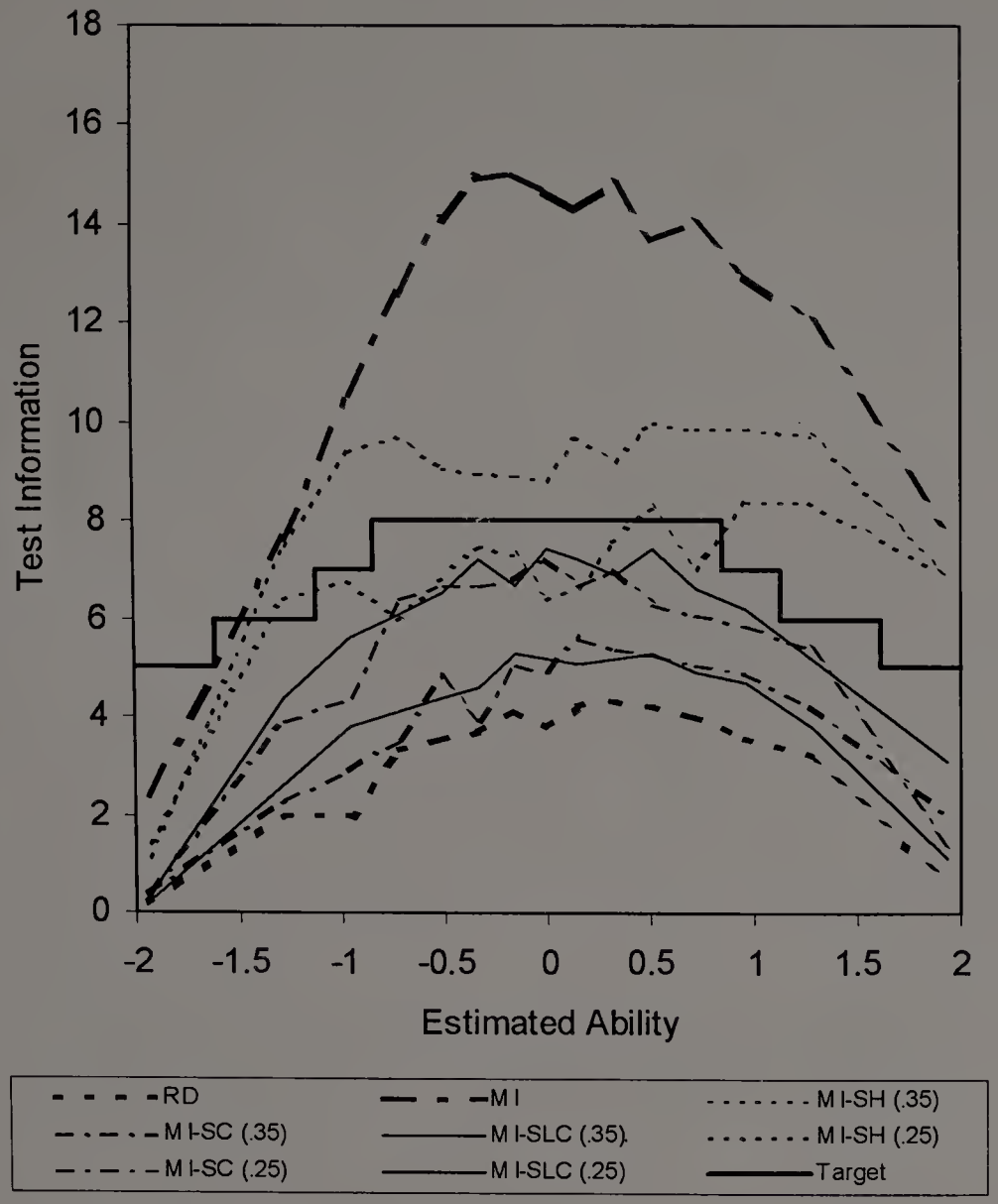
Overall, the results were almost unchanged when the simulations were conducted with the examinee samples that did not match the expected examinee population (average ability increased by .5). Actually, the results were slightly improved, which is most likely explained by the fact that the pools were better suited for these more able than expected samples (Figure 3.1). It is most likely though that the results would have been degraded with for a less able than expected sample.

The extremely low minimum test information values obtained under almost all conditions were troublesome and justified more detailed investigation.

While the overall results facilitated comparisons between test assembly procedures, the conditional results provided more details for the evaluation of each method's merits given the testing situation. To limit the amount of data presented, only the matched sample results were presented. Comparisons between minimum test information and minimum test information targets, standard deviation of test information, proportion of tests providing the required information, and mean test information results computed for each of the 15 ability groups formed based on ability estimates are reported in Figure 4.1 to 4.4.

With the small pool, only the MI and MI-SH (under .35 and .25 unconditional exposure limits could provide sufficiently informative tests. With the larger pool, although measurement was generally poorer for low ability examinees and to a lesser

a) Small item pool



b) Large item pool

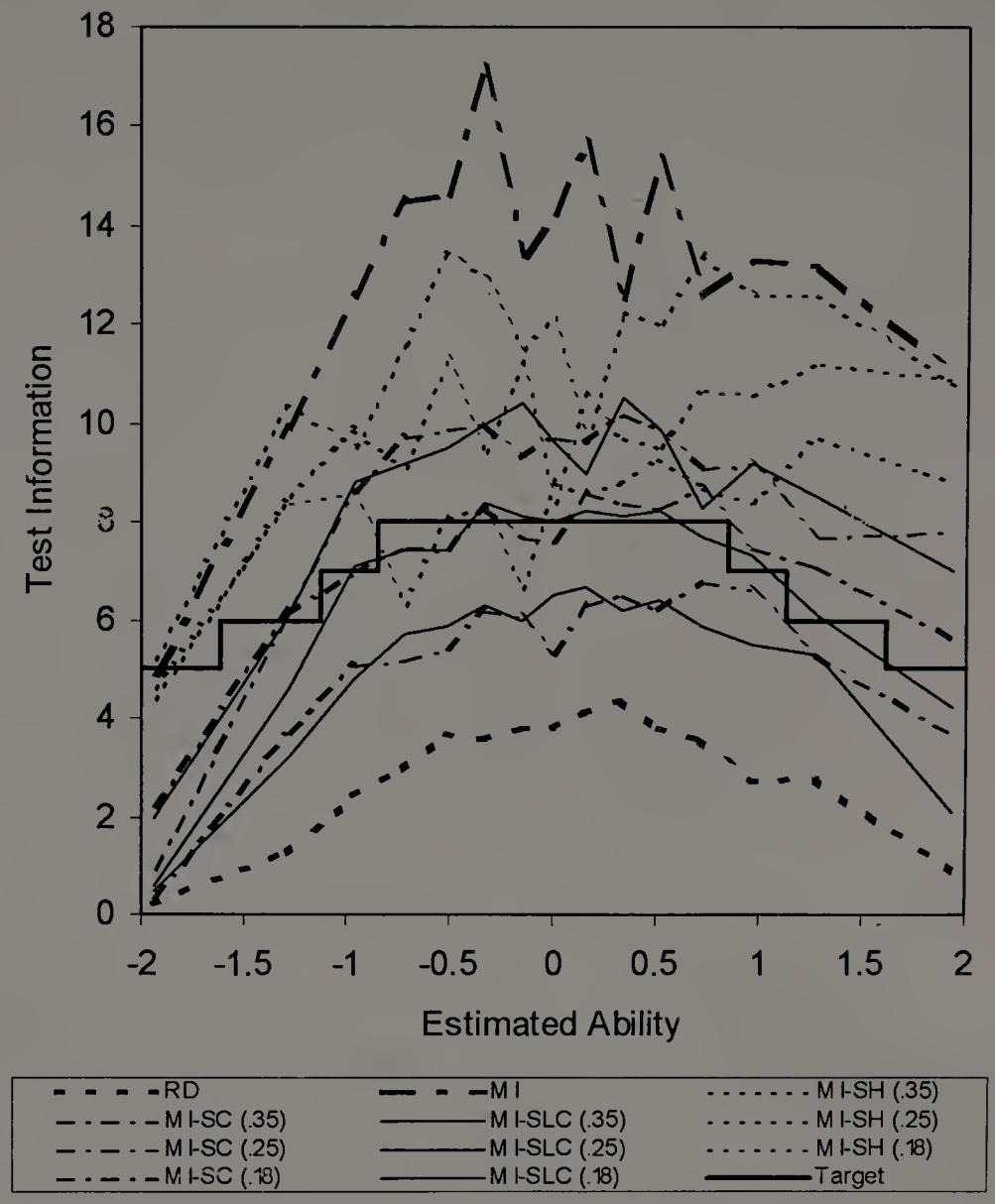
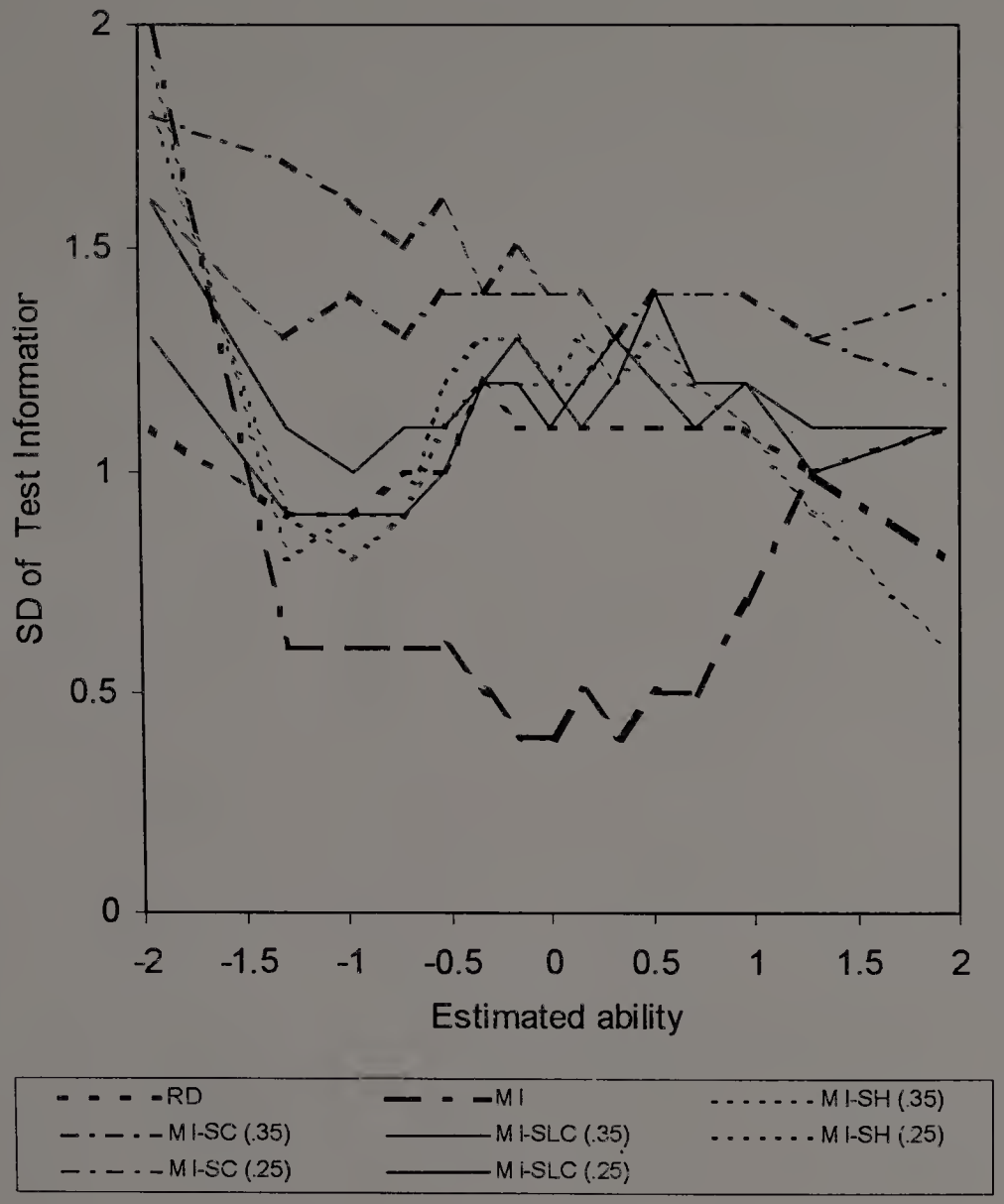


Figure 4.1. Minimum Test Information and Minimum Test Information Targets

a) Small item pool



b) Large item pool

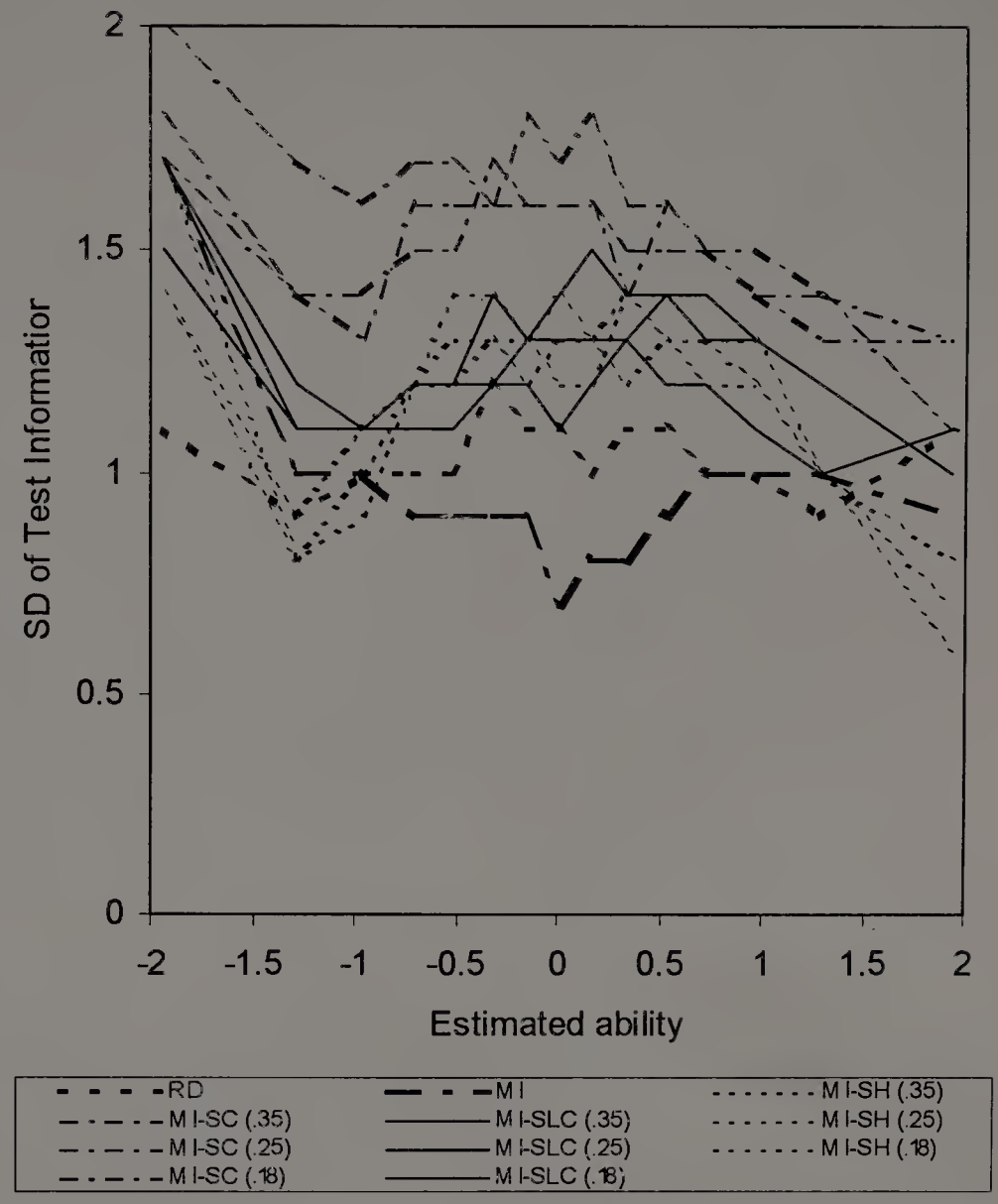
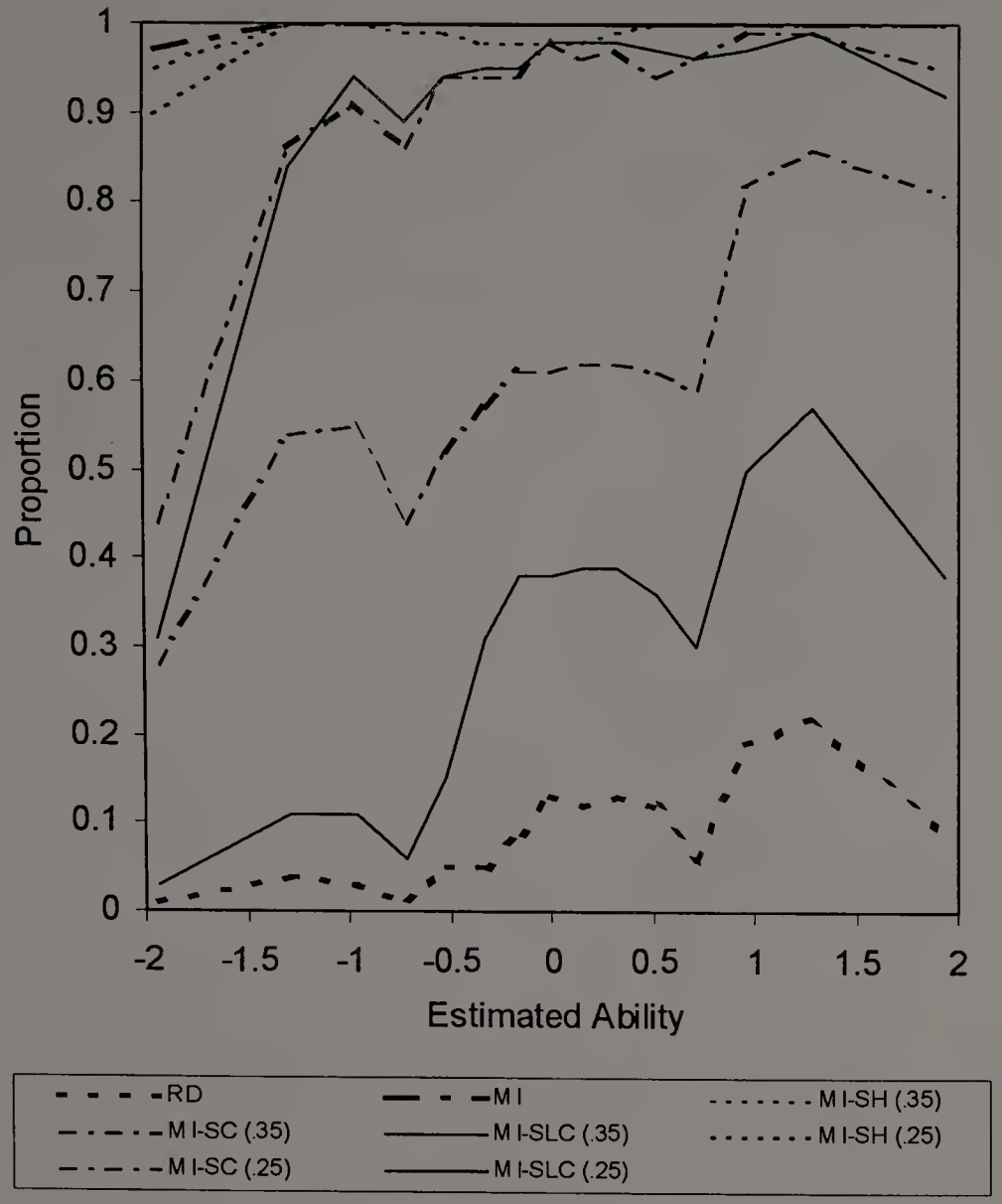


Figure 4.2. Standard Deviation of Test Information

a) Small item pool



b) Large item pool

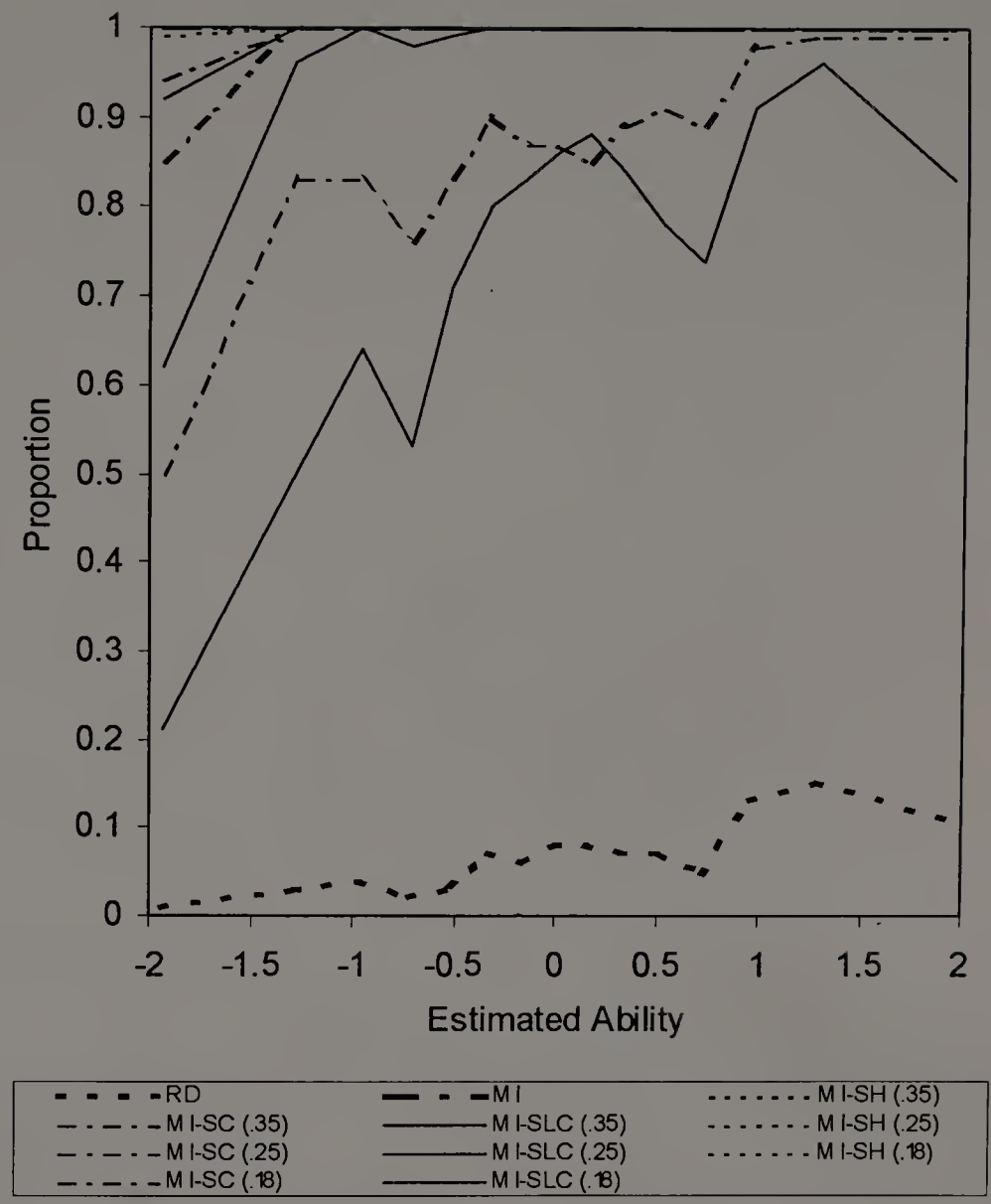
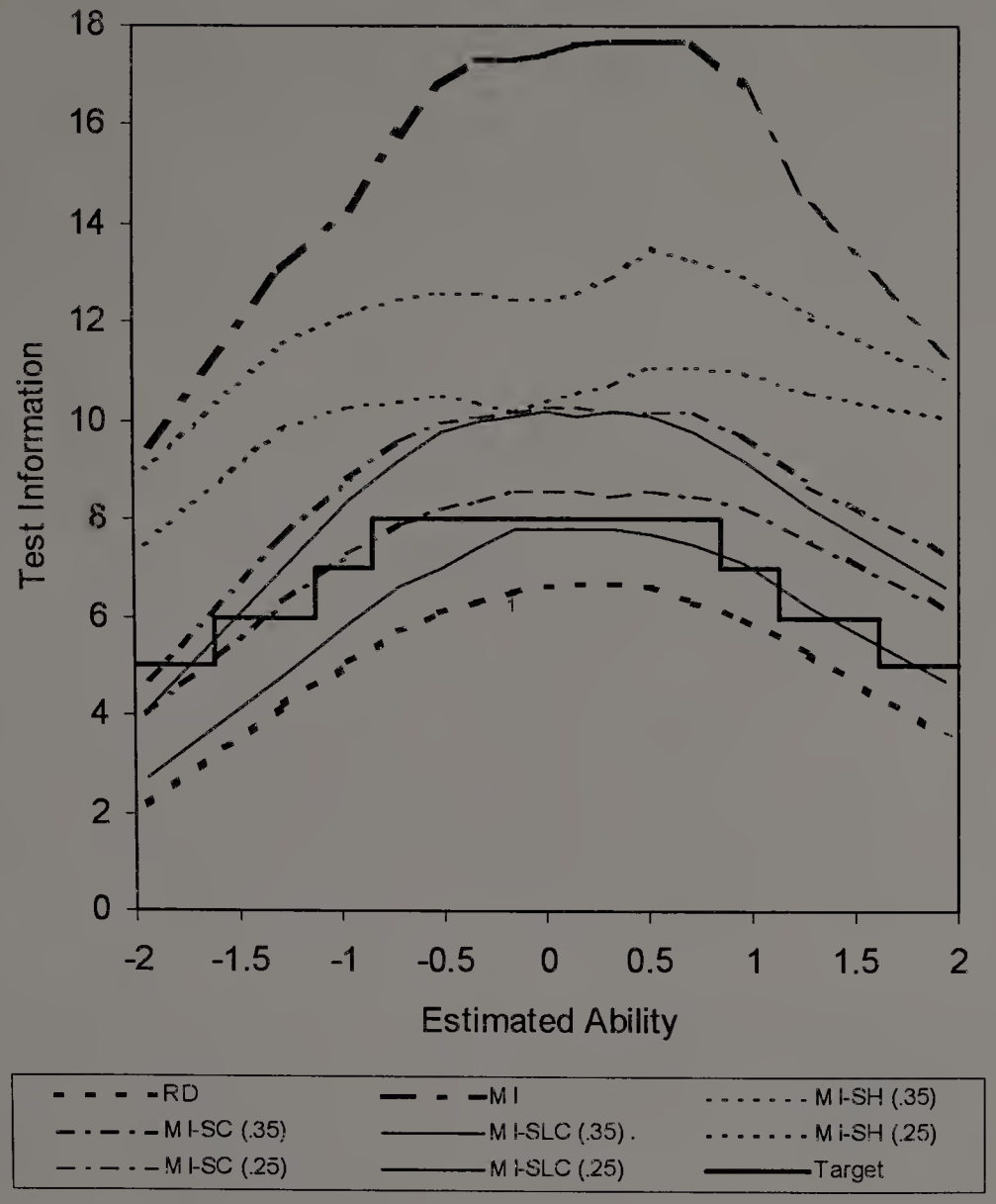


Figure 4.3. Proportion of Tests Providing the Minimum Required Information

a) Small item pool



b) Large item pool

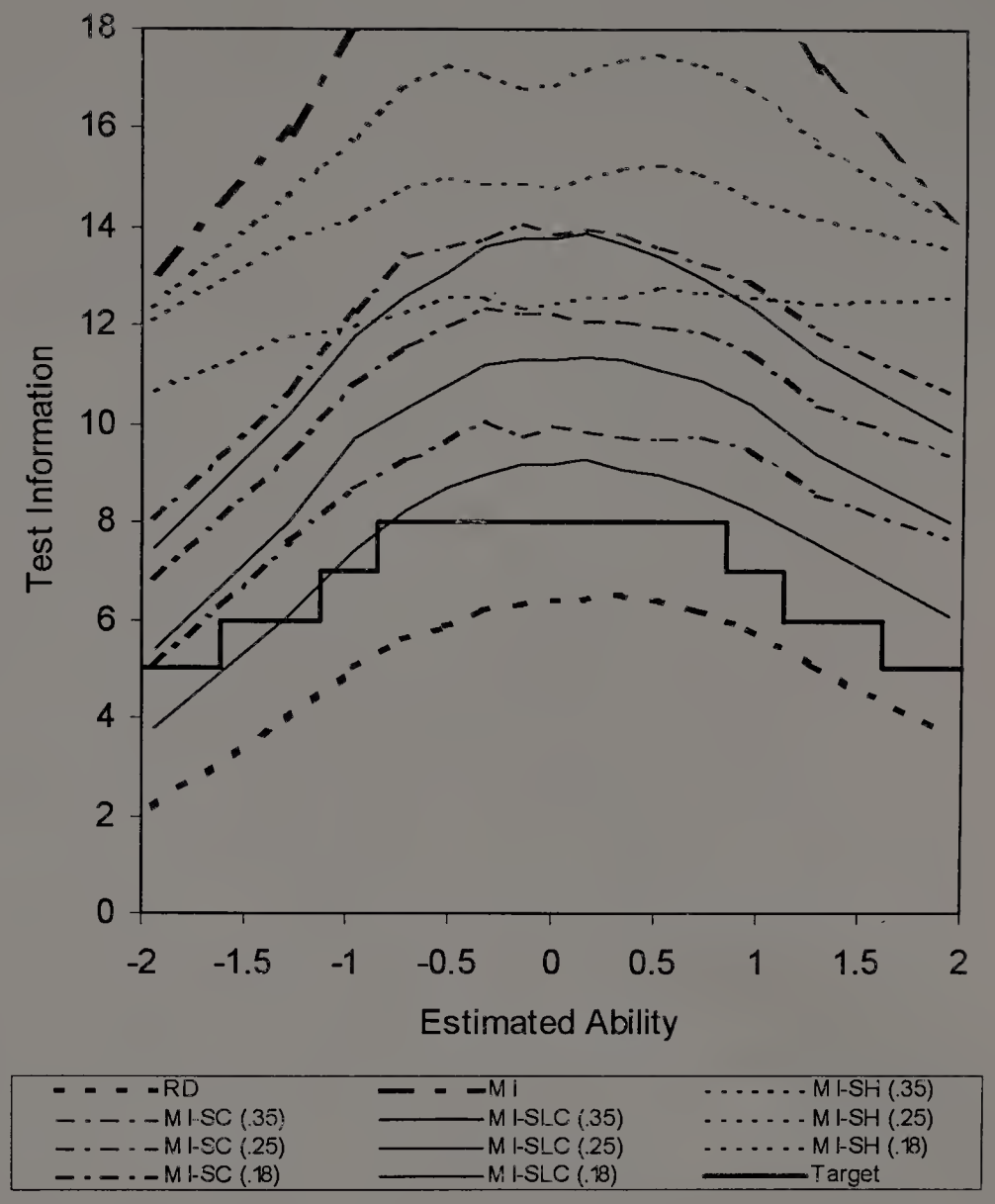


Figure 4.4. Mean Test Information

extent for high ability examinees, performance was substantially improved allowing more severe exposure control. Again, MI and MI-SH (under .35, .25 and .18 unconditional exposure limits), but also MI-SC (under .35, .25 and .18 conditional exposure limits) and MI-SLC (.35 and .25) provided appropriate test information to nearly all examinees.

Also to be noticed are the different levels of test information variability (SD) across procedures. Generally, variability increased with increased exposure control, from about .5 with MI to about 1.5 with the other procedures. More variability was observed at the lower and higher ends of the ability spectrum. Test information variability tended to increase when the larger pool was used, but not by much and around higher average information values.

4.1.2 Measurement Error and Ability Estimation

Observed measurement error (bias and SEM) and ability estimation results (TSEM and ESEM) are presented in Table 4.2. Overall bias was small under all simulation conditions. TSEM as well ESEM were always very close to SEMs, underestimating test precision by at most .2 points indicating that both estimations methods performed well in the testing situation considered. These results confirmed expectations given the comfortable test length used and the good model data fit (since data were simulated as such).

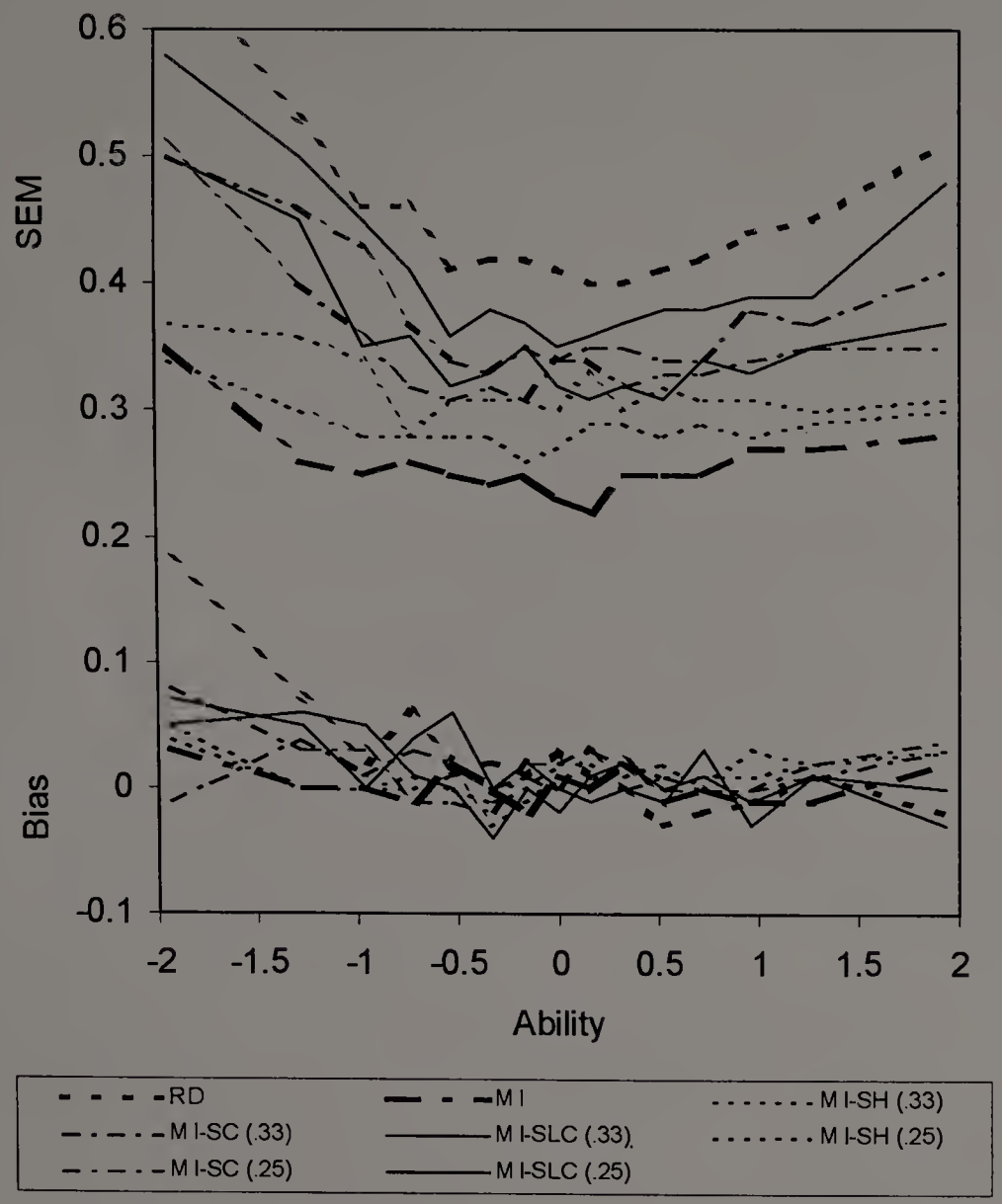
For more details, conditional bias and SEM results with matched sample are presented in Figure 4.5. These results underscore the variability of measurement performance across ability levels. Both bias and SEM increased as abilities were

Table 4.2

Overall Measurement Error Statistics (N=6,000)

Test Assembly Method (Exposure Specifications)	Bias	SEM	TSEM	ESEM
Small Pool (n=200), Expected Sample				
RD	0.03	0.46	0.48	0.47
MI-SH (.35)	0.00	0.29	0.29	0.29
MI-SC (.35)	0.01	0.35	0.35	0.35
MI-SLC(.35)	0.01	0.36	0.36	0.36
MI-SH (.25)	0.00	0.32	0.32	0.32
MI-SC (.25)	0.02	0.39	0.40	0.40
MI-SLC(.25)	0.02	0.41	0.42	0.43
MI	0.00	0.27	0.26	0.26
Larger Pool (n=400), Expected Sample				
RD	0.03	0.46	0.48	0.47
MI-SH (.35)	0.00	0.26	0.25	0.25
MI-SC (.35)	0.01	0.29	0.29	0.29
MI-SLC(.35)	0.00	0.29	0.29	0.29
MI-SH (.25)	0.00	0.27	0.26	0.26
MI-SC (.25)	0.01	0.31	0.31	0.31
MI-SLC(.25)	0.02	0.33	0.33	0.33
MI-SH (.18)	0.00	0.29	0.29	0.29
MI-SC (.18)	0.02	0.34	0.34	0.34
MI-SLC(.18)	0.01	0.37	0.37	0.38
MI	0.00	0.24	0.23	0.24
Small Pool (n=200), Unexpected Sample (+.5 in mean ability)				
MI-SH (.35)	0.01	0.29	0.29	0.29
MI-SC (.35)	0.01	0.34	0.34	0.34
MI-SLC(.35)	0.01	0.35	0.35	0.35
MI-SH (.25)	0.01	0.31	0.31	0.32
MI-SC (.25)	0.01	0.37	0.38	0.38
MI-SLC(.25)	0.00	0.41	0.41	0.41
Larger Pool (n=400), Unexpected Sample (+.5 in mean ability)				
MI-SH (.35)	0.01	0.25	0.25	0.26
MI-SC (.35)	0.01	0.29	0.29	0.29
MI-SLC(.35)	0.00	0.29	0.29	0.29
MI-SH (.25)	0.01	0.27	0.26	0.26
MI-SC (.25)	0.01	0.31	0.31	0.30
MI-SLC(.25)	0.00	0.32	0.32	0.32
MI-SH (.18)	0.01	0.28	0.28	0.28
MI-SC (.18)	0.02	0.34	0.34	0.35
MI-SLC(.18)	0.01	0.36	0.37	0.37

a) Small item pool



b) Large item pool

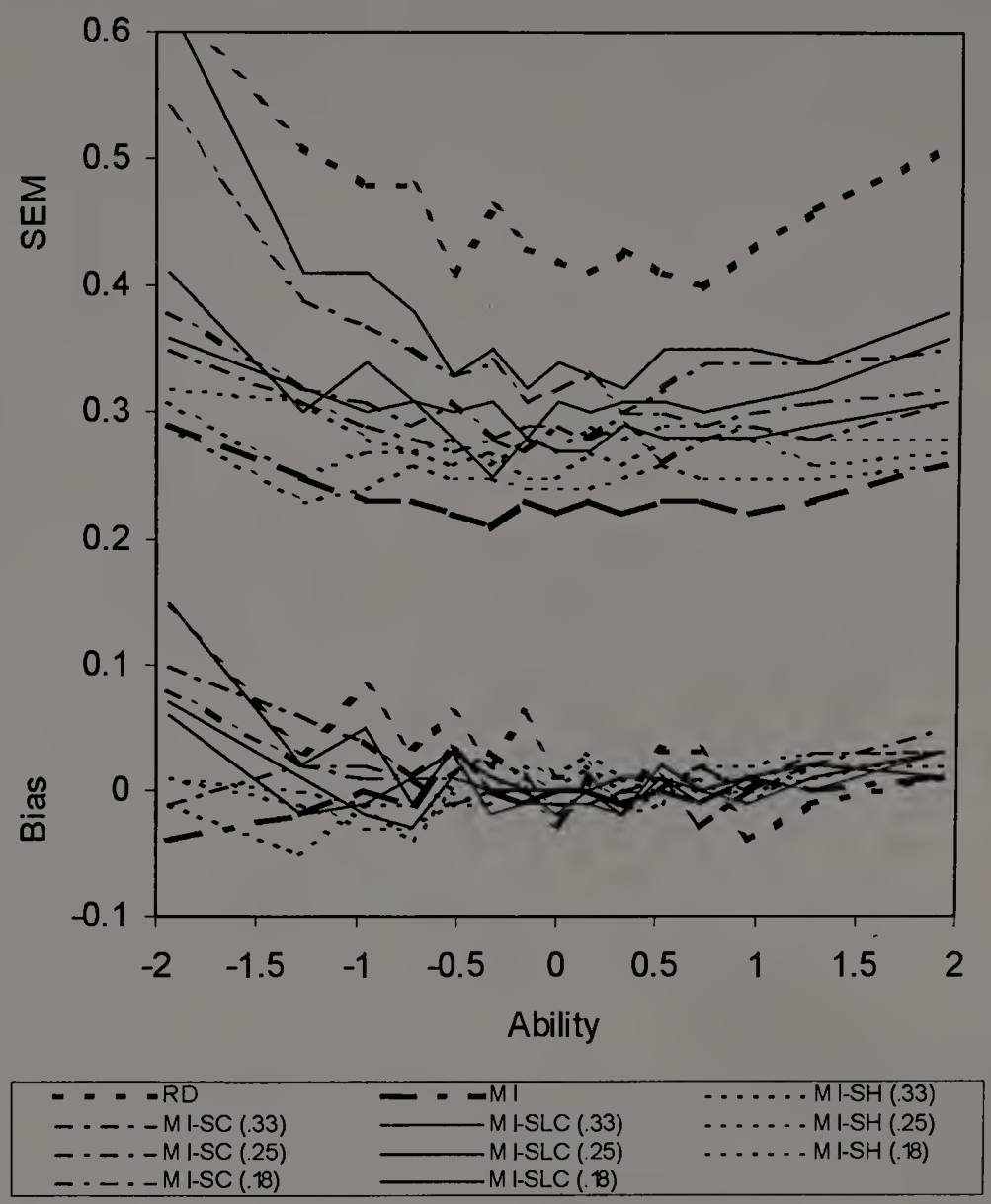


Figure 4.5. Observed Bias and Standard Error of Measurement

further away from average. This was particularly true towards low ability levels and was clearly amplified when more constraints (small pool, severe exposure control) were placed on test assembly.

4.1.3 Summary

With the small pool, only the MI and MI-SH procedures were able to provide the desired minimum level of measurement for all examinees at all ability levels. With the large pool, MI-SC and MI-SLC with the least demanding exposure control (.35) also met the measurement objectives. Although average test information values were still at relatively good levels, more severe exposure control resulted in some examinees being tested with below standard tests. This happened for a few examinees at the lowest ability levels under .25 exposure control limits. Then, under .18 exposure limits, more than 20% of the test administered were below standards at a wide range of ability levels (Figure 4.3).

Clearly, these results make the point that the variability of CAT test information across examinees should not be ignored as good average measurement do not ensure that all examinees are administered good quality tests. As an example, Figure 4.6 shows the least and most informative tests administered to examinees of true ability $\theta = 0.0$, obtained from MI-SLC (.25) with the small pool.

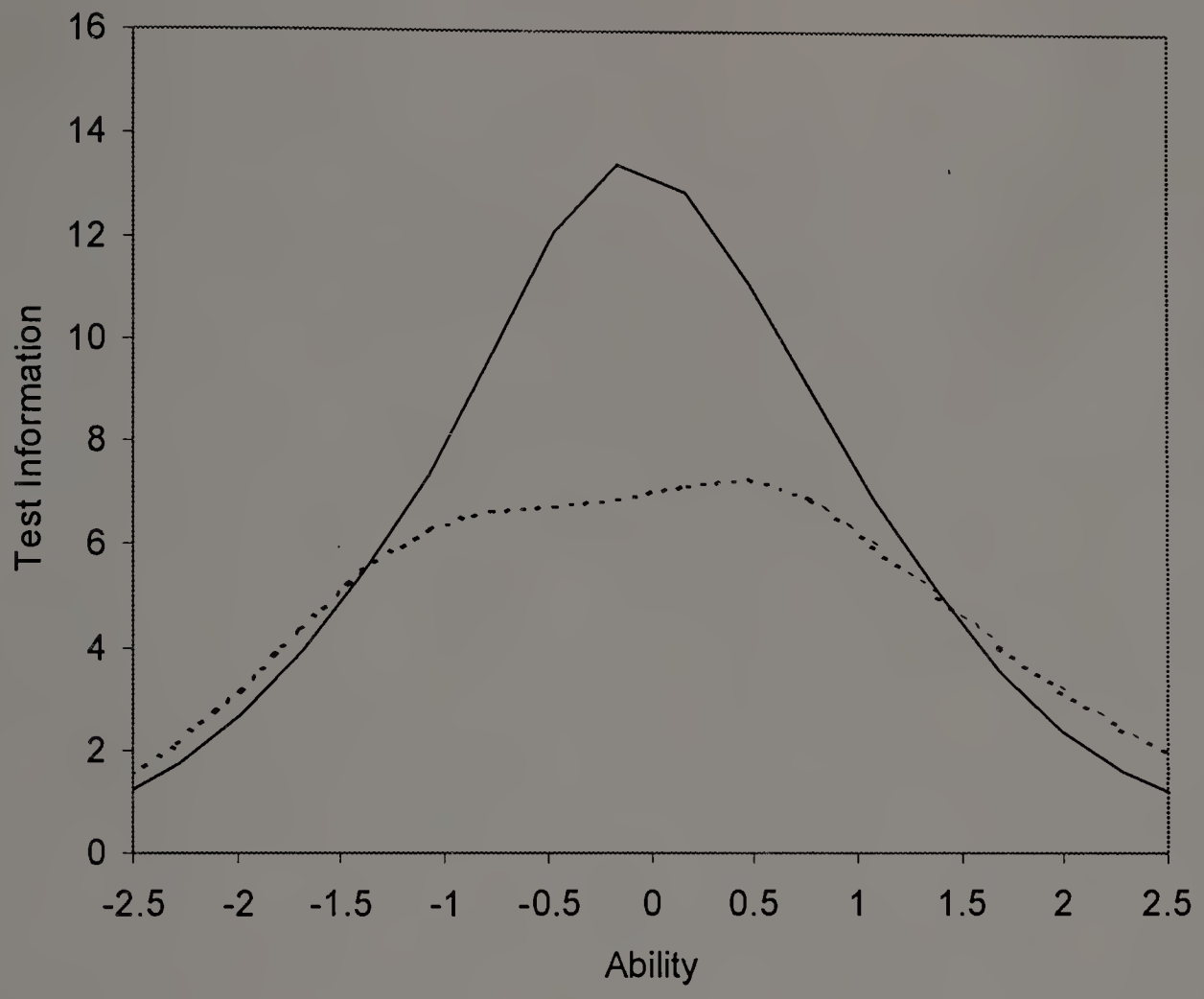


Figure 4.6. Extreme Test Information Functions for Tests Administered to Examinees at the Same Ability Level (0.00)

4.2 Content

Minimum and maximum content specifications were satisfied under most simulation conditions. However, a few (less than 1.5% of the tests) minor deviations started to occur under the most severe exposure control limits (.25 and .18 with the small and larger pools, respectively) and with all three test assembly procedures that used exposure control. Further constraining the test assembly by lowering exposure control limits beyond these values resulted in large proportion of tests to deviate from the ideal content specifications. Therefore, .25 and .18 (small and larger pools, respectively) were considered to be the most constraining settings possible.

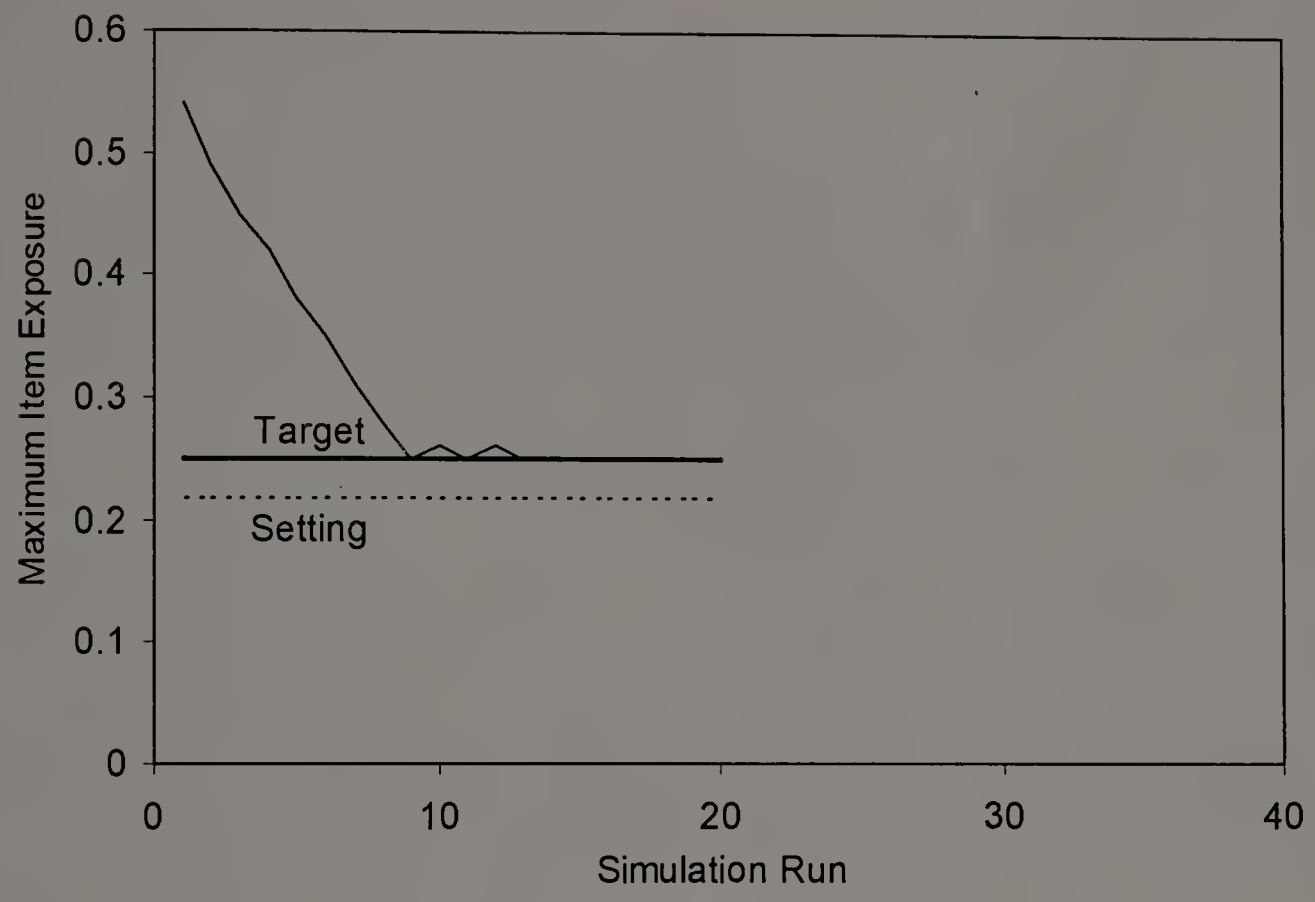
4.3 Security

The results of the preliminary simulation runs necessary for establishing the MI-SH and MI-SLC item exposure parameters are reported first. The security results obtained under all simulation conditions are reported next.

4.3.1 Preliminary Simulations

Iterative simulations of 3,000 tests administrations to examinees representative of the expected population were conducted to determine appropriate SH and SLC item exposure control parameters. As exemplified in Figure 4.7, about 20 and 40 iteration runs were necessary to make sure that parameters converged to stable values with the MI-SH and MI-SLC procedures, respectively. Generally, the observed maximum exposures values obtained in the end exceed set limits (Stocking & Lewis, 1998;

a) MI-SH (.25) Procedure



b) MI-SLC (.25) Procedure (one curve per conditional ability level)

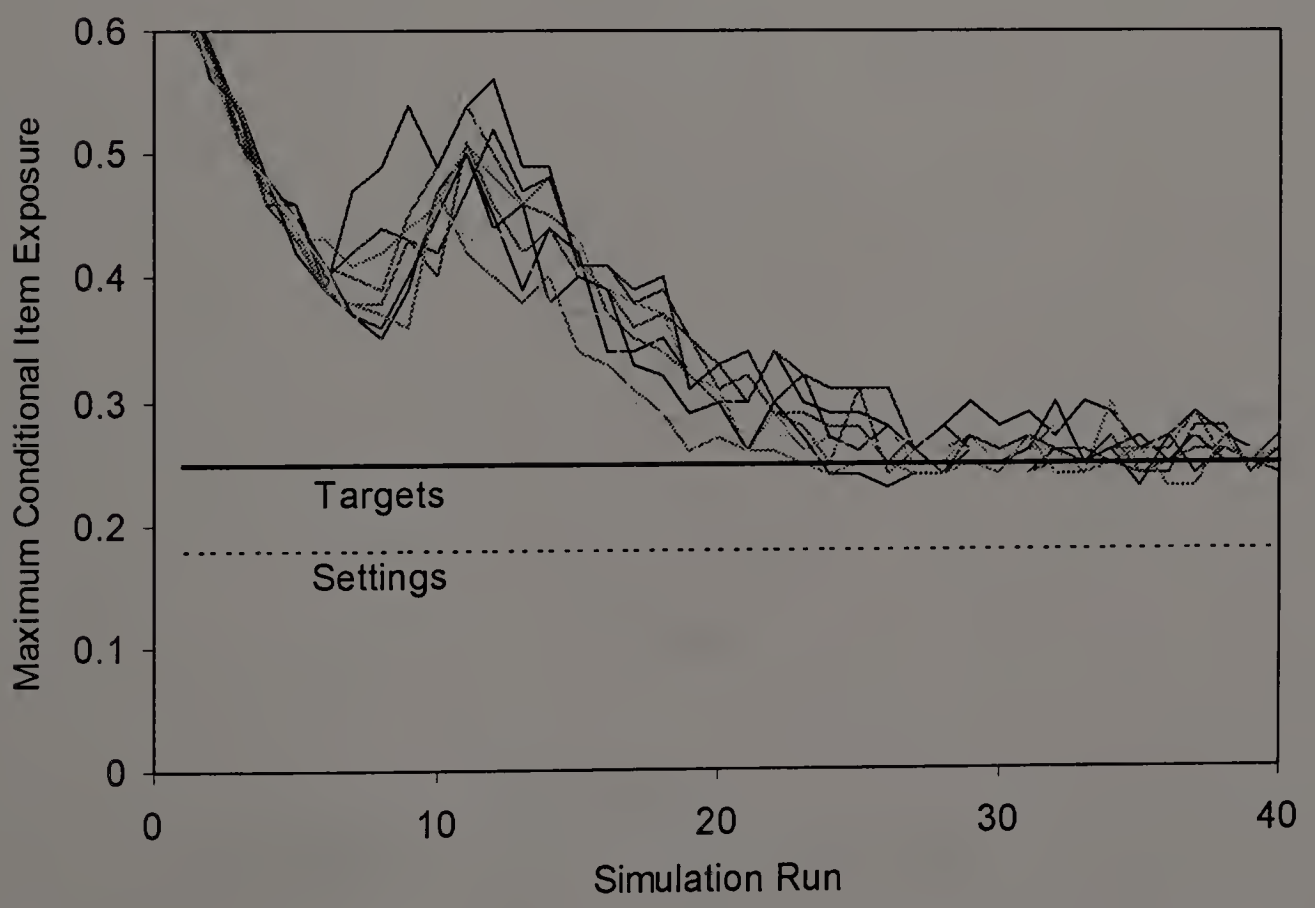


Figure 4.7. Evolution of the Maximum Item Exposure Rates Over Preliminary Simulation Runs

Chang & Twu, 1998), thus repeated iterative simulations, in which settings are progressively adjusted, were necessary to obtain the desired results.

4.3.2 Item Exposure

Detailed results of unconditional and conditional maximum item exposure rates under all simulation conditions are provided in Table 4.3. As expected, RD and MI results were the most extreme. While maximum item exposures could be considered as ideal with the purely random test assembly procedure (similar unconditional and conditional values about .19 and .11 with small and larger pools, respectively), with MI, maximum item exposures were extremely high under all circumstances. The other procedures closely met their unconditional (MI-SH) and conditional (MI-SLC and MI-SC) maximum exposure specifications under all simulation conditions conducted with expected examinee samples. Note that, although MI-SH was not expected to be as effective at reducing conditional exposures as the conditional procedures did, one could have hoped for more substantial reductions. But, even under the most stringent specifications (.25 and .18), it proved to be quite poor at reducing conditional exposures (above .59 and .45 with the small and larger pools, respectively).

The results obtained with the unexpected examinee sample confirmed the robustness of MI-SLC and MI-SC and the lack of robustness of MI-SH procedures to changes in examinee distributions between the simulated examinee sample (generated from the expected population parameters) used to determine the item exposure control parameters and the examinees tested operationally. With MI-SH, unconditional

Table 4.3

Maximum Unconditional and Conditional Item Exposure Rates

Test Assembly Method (Exposure Specifications)	Unconditional	Conditional[a] Mean[b]	Min	Max
Small Pool (n=200), Expected Sample				
RD	0.18	0.19	0.19	0.20
MI-SH (.35)	0.35	0.86	0.71	1.00
MI-SC (.35)	0.32	0.33	0.32	0.34
MI-SLC(.35)	0.33	0.36	0.33	0.38
MI-SH (.25)	0.26	0.72	0.59	0.97
MI-SC (.25)	0.25	0.26	0.25	0.28
MI-SLC(.25)	0.21	0.24	0.22	0.26
MI	0.92	1.00	1.00	1.00
Larger Pool (n=400), Expected Sample				
RD	0.10	0.12	0.11	0.13
MI-SH (.35)	0.36	0.90	0.79	1.00
MI-SC (.35)	0.31	0.33	0.32	0.33
MI-SLC(.35)	0.29	0.35	0.33	0.37
MI-SH (.25)	0.27	0.78	0.64	0.93
MI-SC (.25)	0.23	0.25	0.24	0.26
MI-SLC(.25)	0.21	0.24	0.23	0.27
MI-SH (.18)	0.19	0.61	0.45	0.90
MI-SC (.18)	0.17	0.18	0.17	0.19
MI-SLC(.18)	0.15	0.19	0.17	0.21
MI	0.87	1.00	1.00	1.00
Small Pool (n=200), Unexpected Sample (+.5 in mean ability)				
MI-SH (.35)	0.45	0.86	0.74	1.00
MI-SC (.35)	0.32	0.33	0.32	0.35
MI-SLC(.35)	0.32	0.38	0.34	0.40
MI-SH (.25)	0.34	0.71	0.58	0.98
MI-SC (.25)	0.25	0.26	0.24	0.27
MI-SLC(.25)	0.22	0.25	0.24	0.27
Larger Pool (n=400), Unexpected Sample (+.5 in mean ability)				
MI-SH (.35)	0.48	0.90	0.78	1.00
MI-SC (.35)	0.32	0.33	0.32	0.35
MI-SLC(.35)	0.29	0.35	0.33	0.36
MI-SH (.25)	0.36	0.77	0.60	0.92
MI-SC (.25)	0.24	0.25	0.24	0.26
MI-SLC(.25)	0.20	0.24	0.23	0.27
MI-SH (.18)	0.27	0.60	0.44	0.90
MI-SC (.18)	0.17	0.18	0.17	0.20
MI-SLC(.18)	0.15	0.19	0.17	0.21

[a]: Over 10 conditional ability levels; [b]: Mean of the maximum conditional exposures

maximum exposures increased by .08 to .12 (about 40% increase) as average examinees' ability was .5 points above that of the expected population.

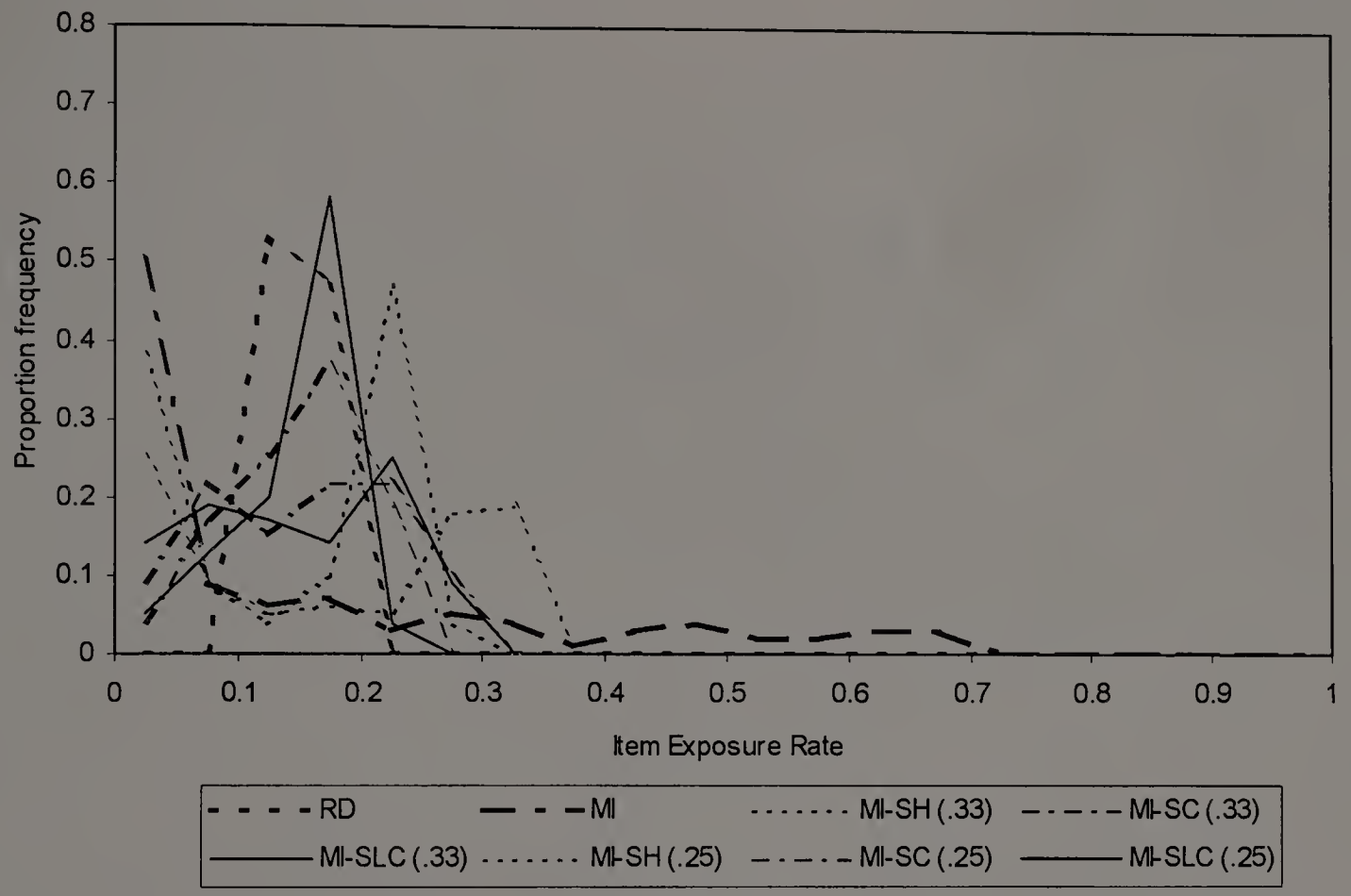
The distributions of unconditional and average conditional item exposures presented in Figure 4.8 and 4.9 provide more information on exposure patterns. Ideally (RD), item exposures should be distributed around their average. Unconditionally, this comes close to being realized when exposure limits are the lowest and conditional exposure control is used. Conditionally, two distinct modes can generally be observed that correspond to groups of items: one includes items that are rarely or not exposed at all, and another one includes items that are exposed at a rate close to the maximum limit. Again, if unconditionally MI-SH results are satisfactory, conditionally, large number of items are rarely or not used at all while exposures are spread over a large range for the other items (Figure 4.9).

4.3.3 Test Overlap

Detailed results of peer-to-peer and test-retest average test overlap rates under all simulation conditions are provided in Table 4.4. Distributions of peer-to-peer and test-retest test overlap rates are provided in Figure 4.10 and 4.11. These results show the effectiveness of the exposure control methods in reducing the number of items in common between any two tests (peer-to-peer overlap) and any two tests given to the same examinee (test-retest overlap).

Peer-to-peer overlap rates were established at .15 and .43 with the RD and MI procedures. With exposure control, results ranged form .27 to .16 (small pool) and .25 to .09 (larger pool). Concerning overlap distributions, the best (concentrated around

a) MI-SH (.25) Procedure



b) MI-SLC (.25) Procedure (one curve per conditional ability level)

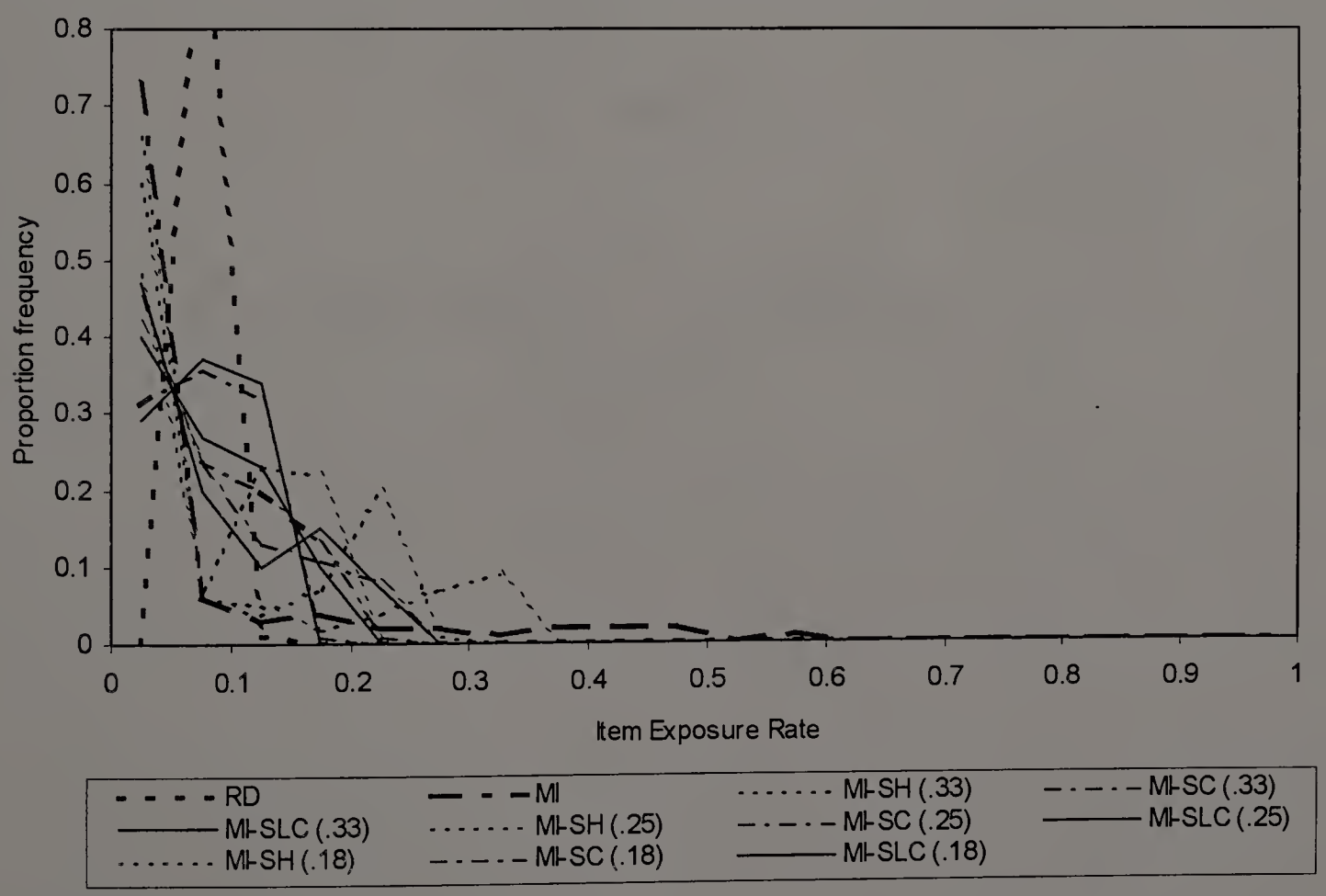
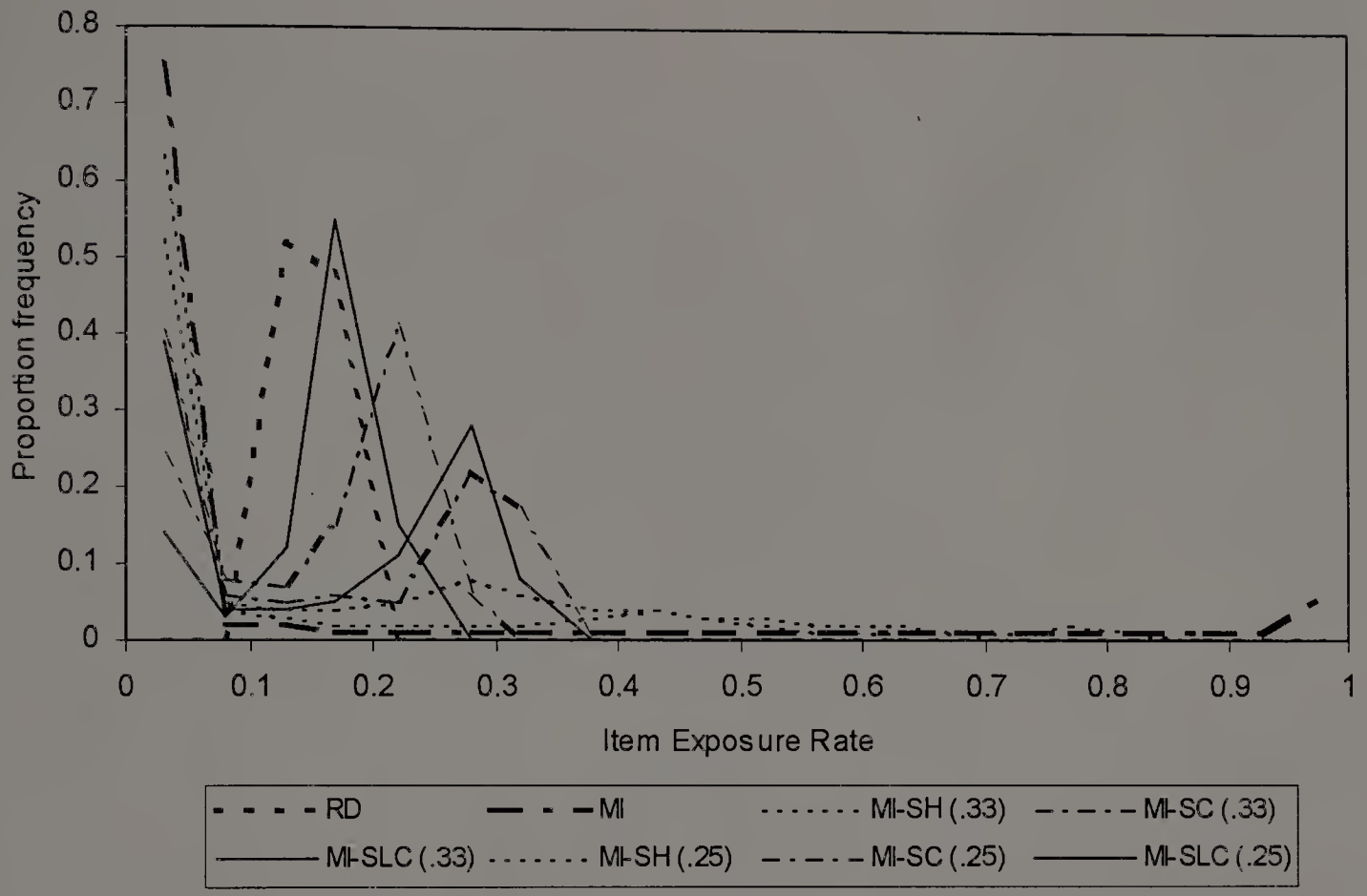


Figure 4.8. Unconditional Item Exposure Distributions

a) MI-SH (.25) Procedure



b) MI-SLC (.25) Procedure (one curve per conditional ability level)

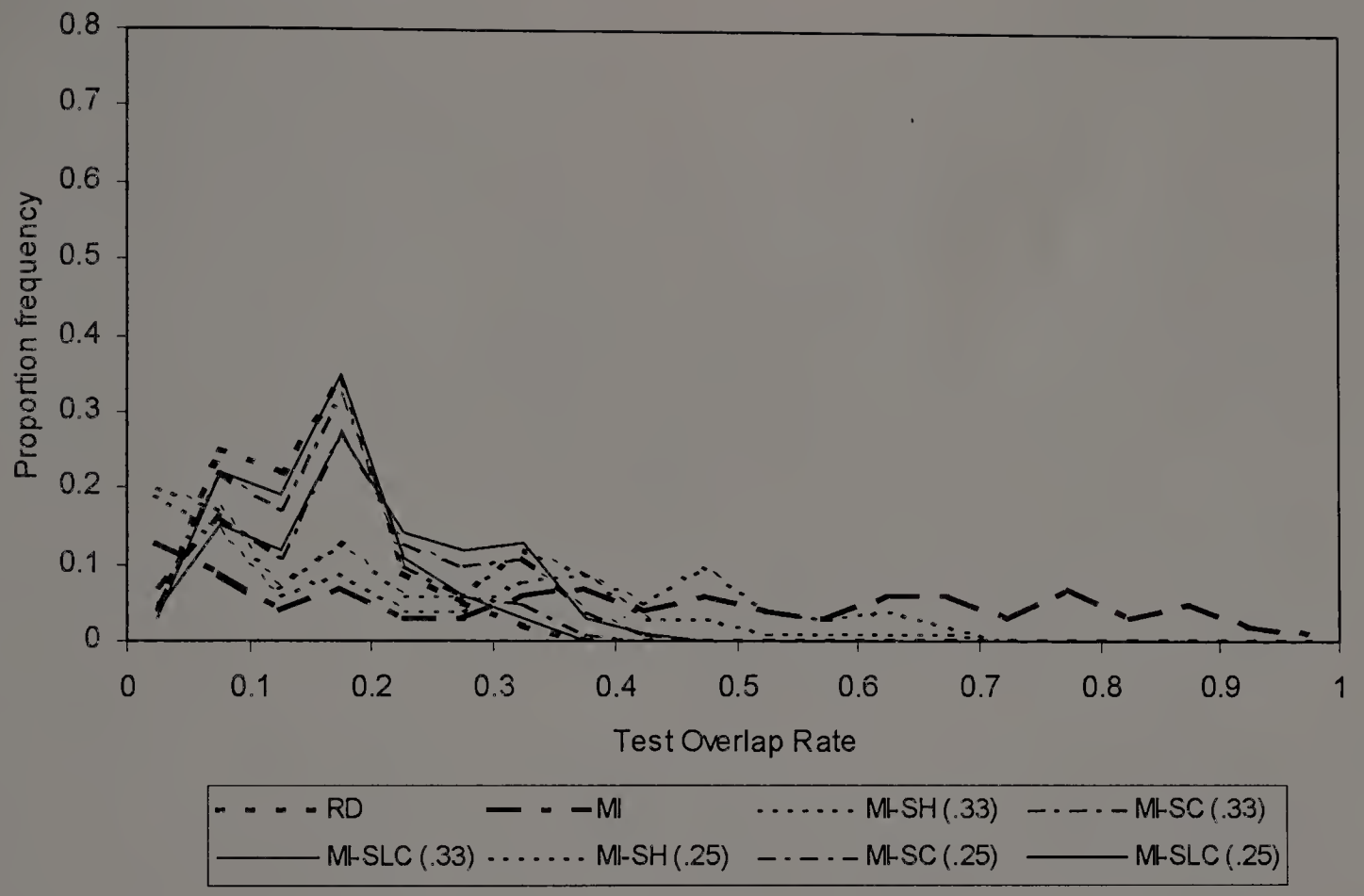
Figure 4.9. Conditional Item Exposure Distributions

Table 4.4

Average Peer-to-peer and Test-retest Overlap Exposure Rates

Test Assembly Method (Exposure Specifications)	Test (Peer-to-peer)	Test-retest Mean	Test-retest Min	Test-retest Max
Small Pool (n=200), Expected Sample				
RD	0.15	0.15	0.15	0.15
MI-SH (.35)	0.27	0.52	0.45	0.66
MI-SC (.35)	0.19	0.26	0.26	0.28
MI-SLC(.35)	0.19	0.26	0.25	0.27
MI-SH (.25)	0.21	0.39	0.31	0.52
MI-SC (.25)	0.17	0.20	0.19	0.23
MI-SLC(.25)	0.16	0.18	0.17	0.18
MI	0.43	0.78	0.70	0.83
Larger Pool (n=400), Expected Sample				
RD	0.08	0.08	0.08	0.08
MI-SH (.35)	0.25	0.52	0.46	0.61
MI-SC (.35)	0.15	0.25	0.25	0.27
MI-SLC(.35)	0.15	0.24	0.24	0.25
MI-SH (.25)	0.20	0.42	0.35	0.57
MI-SC (.25)	0.12	0.20	0.19	0.21
MI-SLC(.25)	0.11	0.17	0.16	0.17
MI-SH (.18)	0.14	0.31	0.24	0.45
MI-SC (.18)	0.10	0.14	0.13	0.15
MI-SLC(.18)	0.09	0.12	0.12	0.12
MI	0.38	0.74	0.64	0.81
Small Pool (n=200), Unexpected Sample (+.5 in mean ability)				
MI-SH (.35)	0.28	0.52	0.45	0.65
MI-SC (.35)	0.19	0.27	0.26	0.29
MI-SLC(.35)	0.19	0.26	0.25	0.27
MI-SH (.25)	0.22	0.39	0.32	0.51
MI-SC (.25)	0.17	0.20	0.19	0.22
MI-SLC(.25)	0.16	0.18	0.17	0.18
Larger Pool (n=400), Unexpected Sample (+.5 in mean ability)				
MI-SH (.35)	0.27	0.52	0.46	0.61
MI-SC (.35)	0.15	0.26	0.25	0.27
MI-SLC(.35)	0.15	0.24	0.23	0.25
MI-SH (.25)	0.21	0.42	0.35	0.57
MI-SC (.25)	0.12	0.20	0.19	0.21
MI-SLC(.25)	0.11	0.17	0.16	0.17
MI-SH (.18)	0.16	0.31	0.24	0.44
MI-SC (.18)	0.10	0.14	0.13	0.15
MI-SLC(.18)	0.09	0.12	0.12	0.12

a) MI-SH (.25) Procedure



b) MI-SLC (.25) Procedure (one curve per conditional ability level)

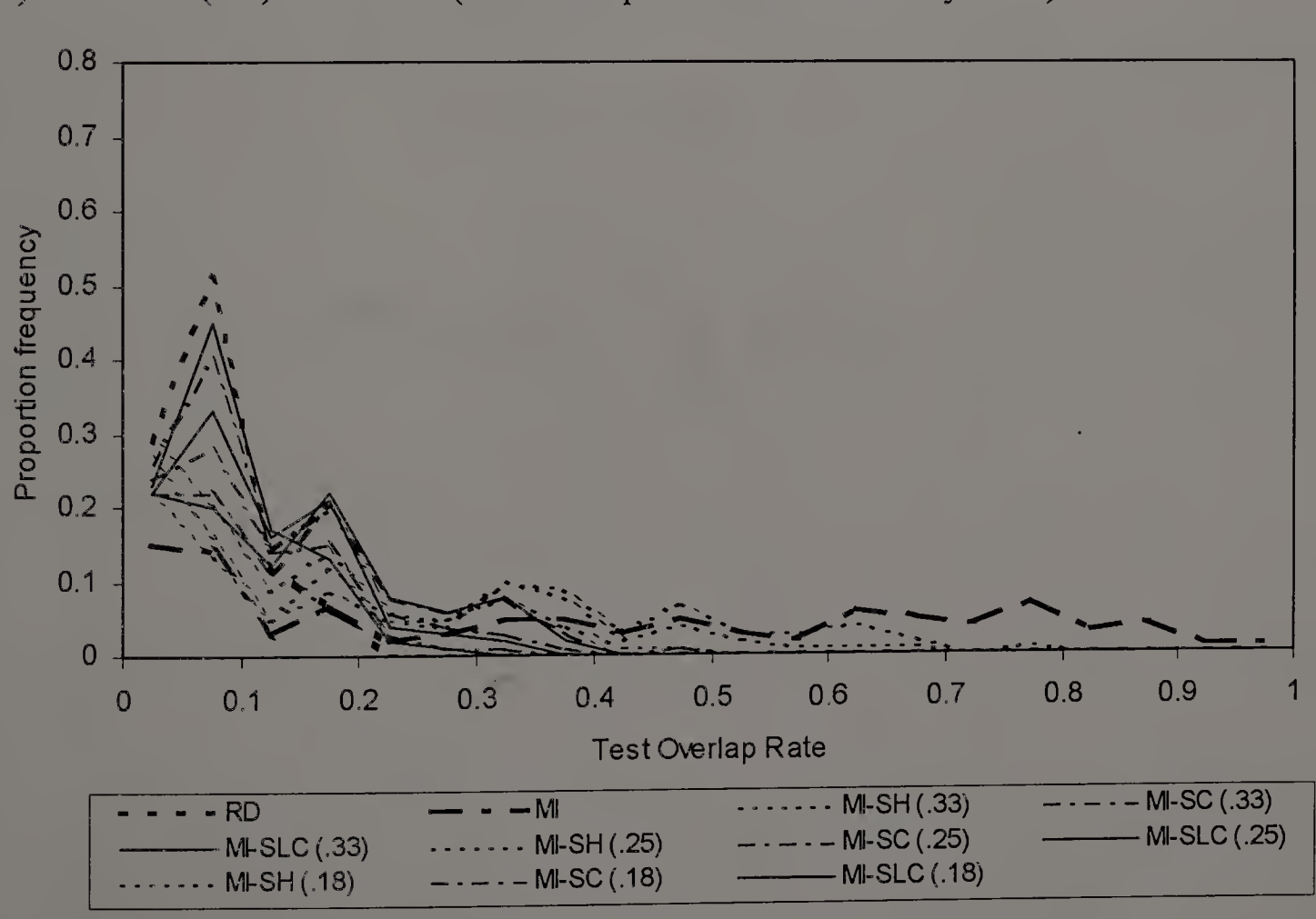
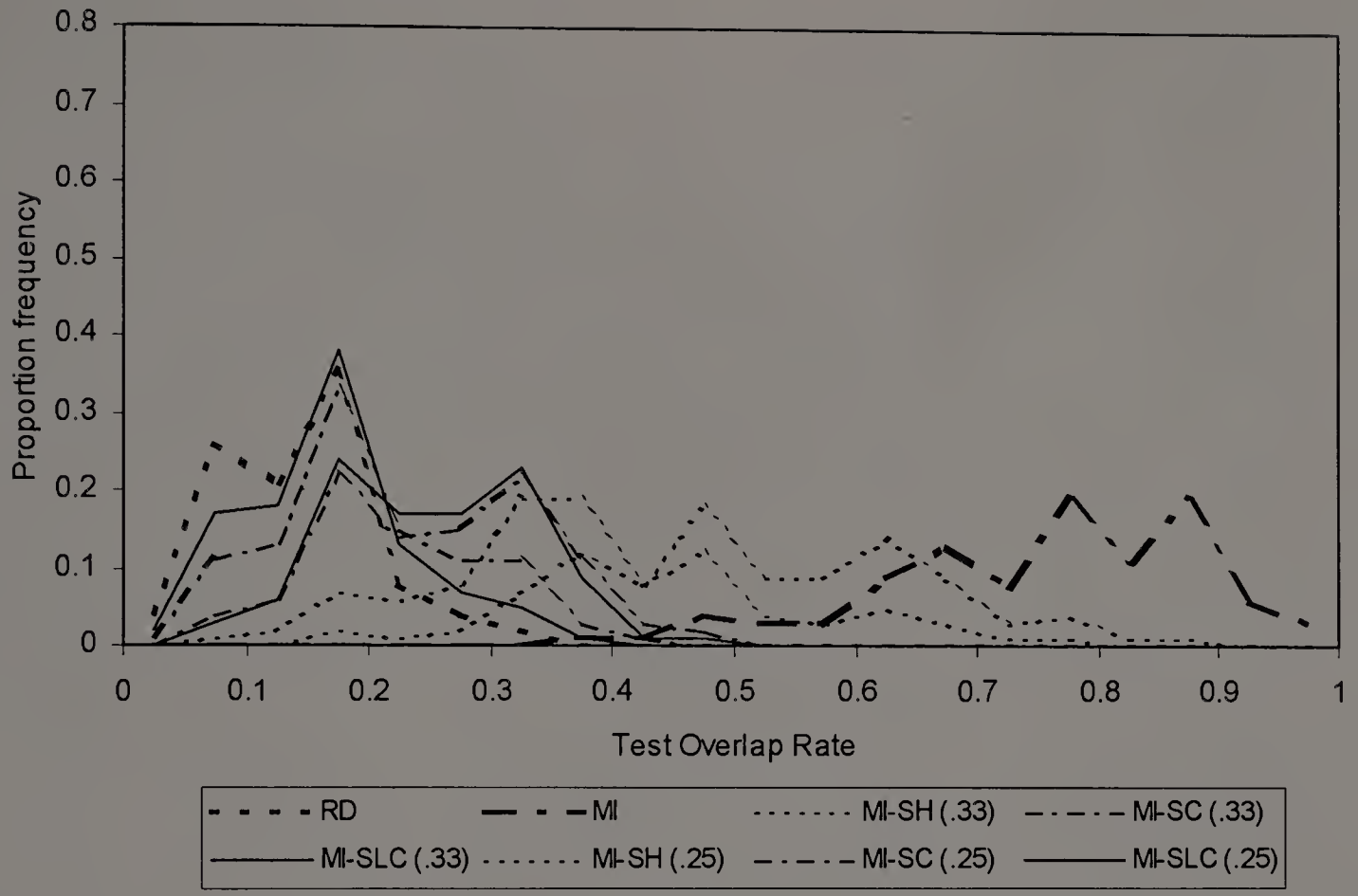


Figure 4.10. Peer-to-peer Overlap Distributions

a) MI-SH (.25) Procedure



b) MI-SLC (.25) Procedure (one curve per conditional ability level)

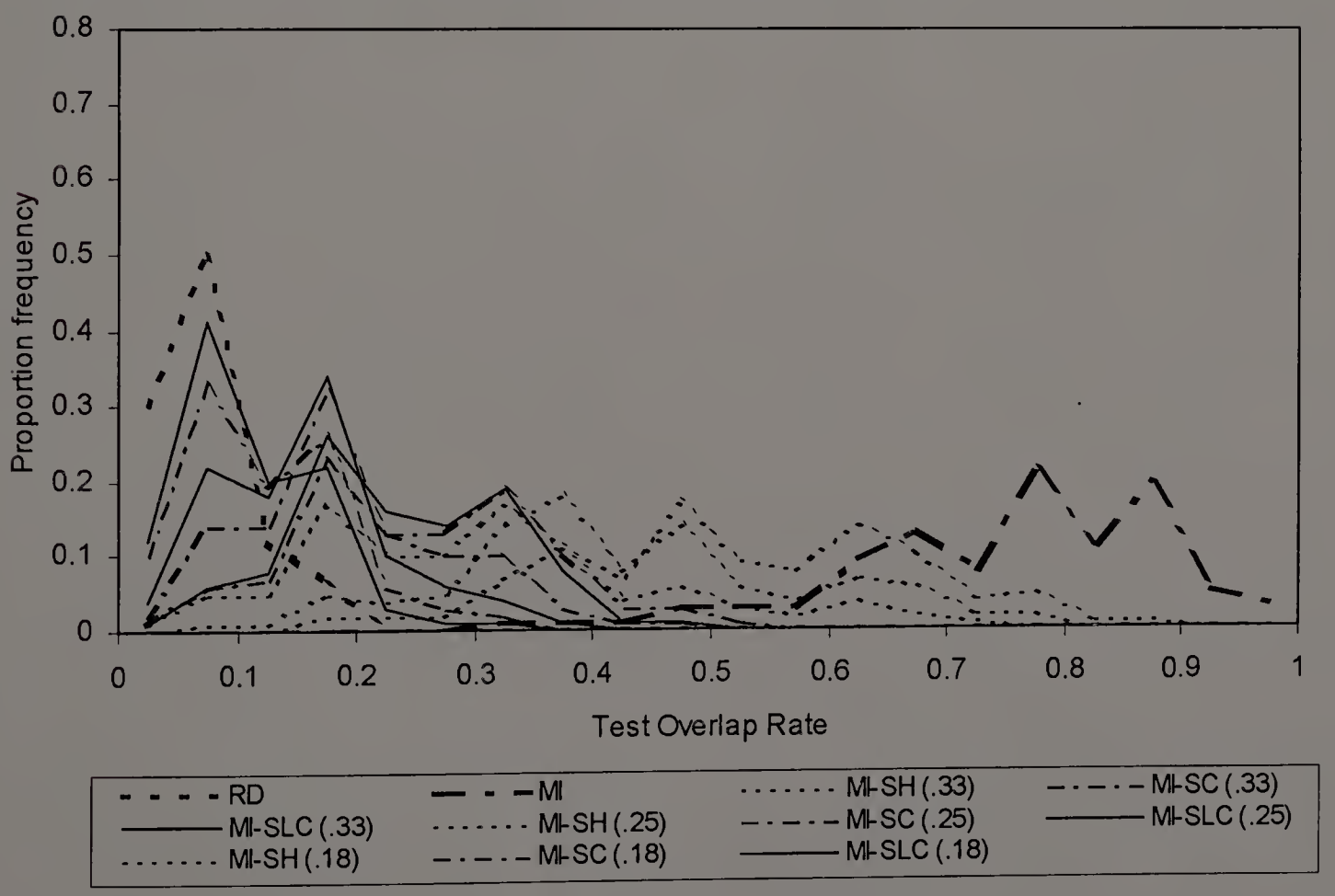


Figure 4.11. Test-retest Overlap Distributions

the mean) and worst (spread with noticeable proportional frequencies towards high overlap rates) patterns are exemplified by those obtained with the RD and MI procedures. Again the best results, by far, were obtained with the MI-SC and MI-SLC procedures under the most severe exposure control limits.

4.3.4 Summary

In term of security, important differences exist between test assembly with and without exposure control, and between unconditional and conditional exposure control. The major problem of the maximum information strategy, demonstrated many times and repeated here, is that very few of the items available are used. This results in administering tests that have a very large proportion of items in common, especially for examinees of similar ability (70% and more). With unconditional exposure control (MI-SH), overall item exposure and peer-to-peer test overlap can be controlled precisely and reduced to low values. But still, tests administered to examinees of similar ability have large numbers of items in common (45% and more). Therefore, MI-SH may be a good approach for low-stakes testing, as it does provide some security and to a degree make use of more of the items available. Clearly, a conditional exposure control procedure is necessary to ensure the level of security needed in high-stakes testing conditions. Both, MI-SC and MI-SLC demonstrated a similar capability to provide high security to test administration.

4.4 Efficiency

The relative efficiency of each test assembly procedure was judged based on the extent to which it does (or does not) satisfy measurement and content specifications, provide desirable security and makes the best use of the available items. Table 4.5 provides overall evaluations of testing efficiency under each simulation condition.

As expected, the random item selection procedure provided the highest testing security but was unable to provide acceptable measurement. The maximum information procedure provided the highest measurement properties but was unable to provide any security. Consequently, given our testing objectives and their operational specifications, both procedures were dismissed as unacceptable and thus inefficient.

The MI-SH procedure satisfied both measurement and content specifications but provided enough testing security only for low to moderate-stakes testing situations, with both small and larger pools. Increasing exposure control improved the χ^2 pool usage index from 23 to 11 and from 69 to 27 with the small and larger pools, respectively. However it should be noted that MI-SH did not reach the desired level of exposure control when the distribution of examinee tested differed from that of the simulation sample used in determining the exposure control parameters (unexpected sample condition).

The MI-SC and MI-SLC procedures both satisfied measurement and content specifications and provided high security appropriate for high-stakes testing situations. However, the small pool was not sufficient to support testing with either of these procedures. As with MI-SH, it was observed that increased exposure control resulted in

Table 4.5

Overall Evaluation of Testing Efficiency

Test Assembly Method (Exposure Specifications)	Measurement[a]	Content[a]	Security	χ^2 Pool Usage Index
Small Pool (n=200), Expected Sample				
RD	No	Yes	Best	0
MI-SH (.35)	Yes	Yes	Unconditional	23
MI-SC (.35)	No	Yes	Conditional	8
MI-SLC(.35)	No	Yes	Conditional	9
MI-SH (.25)	Yes	Yes	Unconditional	11
MI-SC (.25)	No	Yes	Conditional	4
MI-SLC(.25)	No	Yes	Conditional	3
MI	Yes	Yes	Worst	55
Larger Pool (n=400), Expected Sample				
RD	No	Yes	Best	1
MI-SH (.35)	Yes	Yes	Unconditional	69
MI-SC (.35)	Yes	Yes	Conditional	29
MI-SLC(.35)	Yes	Yes	Conditional	28
MI-SH (.25)	Yes	Yes	Unconditional	48
MI-SC (.25)	Yes	Yes	Conditional	19
MI-SLC(.25)	Yes	Yes	Conditional	15
MI-SH (.18)	Yes	Yes	Unconditional	27
MI-SC (.18)	No	Yes	Conditional	9
MI-SLC(.18)	No	Yes	Conditional	8
MI	Yes	Yes	Worst	121
Small Pool (n=200), Unexpected Sample (+.5 in mean ability)				
MI-SH (.35)	Yes	Yes	Unconditional	27
MI-SC (.35)	No	Yes	Conditional	8
MI-SLC(.35)	No	Yes	Conditional	9
MI-SH (.25)	Yes	Yes	Unconditional	14
MI-SC (.25)	No	Yes	Conditional	4
MI-SLC(.25)	No	Yes	Conditional	2
Larger Pool (n=400), Unexpected Sample (+.5 in mean ability)				
MI-SH (.35)	Yes	Yes	Unconditional	77
MI-SC (.35)	Yes	Yes	Conditional	31
MI-SLC(.35)	Yes	Yes	Conditional	30
MI-SH (.25)	Yes	Yes	Unconditional	53
MI-SC (.25)	Yes	Yes	Conditional	19
MI-SLC(.25)	Yes	Yes	Conditional	15
MI-SH (.18)	Yes	Yes	Unconditional	32
MI-SC (.18)	No	Yes	Conditional	9
MI-SLC(.18)	No	Yes	Conditional	8

[a] Satisfies (Yes or No) specifications;

improved testing efficiency as measured by the χ^2 pool usage index (from 29 to 8, with .35 to .18 conditional exposure limits and the larger item pool).

As indicated earlier, the two designs—simultaneously using either $2p$ small independent pools (N=200) or p larger independent pools (N=400)—are directly comparable in terms overall item usage as measured by Chang and Ying's index. However, for overall comparisons, exposures should be adjusted to take the difference in the number of pools that can be used simultaneously in account. In our case, small pool exposures should be divided by two for comparisons with the larger pool exposures. Consequently, pool usage results between the small pool under .35 exposure control and the larger pool under .18 exposure control can be compared. Interestingly enough, these results are almost identical (slightly better with the small pool), suggesting that, provided enough items are made available, using larger pools than necessary does not necessary improve testing efficiency. Efficiency may be improved by finding the optimum pool size given the testing situation at hand.

CHAPTER 5

SUMMARY AND CONCLUSIONS

In this simulation study the development and thorough evaluation of five CAT test assembly procedures were investigated in conditions representative of typical on-demand testing situation. Important testing objectives where identified, criteria for their evaluation were specified, and comparative results across alternative test assembly procedures were provided. The high-stakes testing situation simulated aimed at providing informative, content balanced (6 content areas) and secure 30-item tests to examinees, given the availability of either 200 or 400 item pools. In addition to pool size, the degree of exposure control and the match between the expected examinee ability distribution used to develop test assembly procedures and the "real" examinee ability distribution were manipulated.

Recognizing the fact that, with on-demand CAT, examinees may not be provided the same opportunity to demonstrate their ability, minimum test information targets set at different ability levels were used as the main criteria for the evaluation of measurement objectives. Minimum and maximum number of items per test content attribute were used for the evaluation of content objectives. Unconditional and conditional item exposure and test overlap rates were used for the evaluation of security objectives. Finally, for any test assembly procedure, overall performance was evaluated based on the most efficient use of the items available and the satisfaction of all measurement, content, and security.

Among the five test assembly procedures investigated, the random (RD) and maximum information (MI) procedures provided measures of the best and worst obtainable tests with respect to measurement and security objectives. They also illustrated the need for compromise between testing objectives, with RD resulting in very secured but poorly informative tests on the one hand, and MI resulting in very informative but totally non-secured tests on the other end. Therefore the capability of three other alternative procedures combining the maximum information item selection strategy with some form of exposure control developed to efficiently realize desirable compromise between testing objectives was investigated in detail. To handle the content specifications and, to a certain degree, optimize the multiple testing objectives, the test assembly procedures were modeled with the weighted deviation model and solved using simple heuristics.

The popular MI-SH procedure, using Sympson & Hetter unconditional exposure control, demonstrated only moderate levels of testing security and proved less efficient than the other two procedures using conditional exposure control (given that enough items were available). Although the MI-SH procedure required preliminary simulations to determine its associated item exposure control parameters, these were relatively easy to conduct. However, it was not robust to violations of the distributional assumptions that are needed in order to execute these preliminary simulations, thus weakening its potential to effectively control item exposure.

The more recent MI-SLC procedure, using Stocking & Lewis conditional exposure control, demonstrated high levels of testing security and proved to be very efficient, given that enough items were made available. However, the procedure

requires very extensive preliminary simulations to determine its operational matrix of conditional item exposure control parameters. The results obtained with the MI-SC procedure, using so-called stochastic conditional exposure control, were very similar to that of the MI-SLC procedure. The advantages of MI-SC over MI-SLC are that it does not require extensive preliminary simulations for the determination of the operational item conditional exposure control parameters, and it allows for instant item removal or replacement from the pool (in case a flaw is discovered) without requiring new preliminary simulations. Both MI-SLC and MI-SC were shown to provide the best compromise between all competing testing objectives in high-stakes situations and also maintained high test security over relatively large changes in the examinees' ability distribution.

Clearly, the availability of large enough item pools is essential for any test assembly procedure to satisfy measurement, content and security objectives. More effective exposure control required larger pool size but also resulted in better test efficiency. However it was also found that for any of the three procedures with exposure control, little efficiency was to be gained by increasing pool size beyond its minimum value. In fact, it appears that although increasing pool size does allow more items to be used more frequently, in proportion, the number of rarely used items increases even more. Rather than increasing pool size, a better approach may be to generate more pools to be used simultaneously rather than fewer large pools.

The variability of test information across examinees at any ability level was found to be large for all procedures and under all simulation conditions. Consequently, it should be taken into account when measurement specifications are set to avoid some

of the examinees being administered substandard tests. The approach adopted here was to require minimum test information targets to be met. In this way, only the procedures capable of consistently producing minimally informative tests could be accepted. Again, the availability of enough items was crucial for the procedures to satisfy this requirement. Although increased exposure control did not appear to increase the variability of test information, it did decrease average test information which lead to violations of the minimum test information requirements.

5.1 Limitations of the Study

As with any simulation study there are questions about the generalizability of the findings to real testing situations. In this study, care was taken to ensure that the simulated conditions where realistic. The IRT model used to generate and analyze data is one that many testing agencies have found appropriate for item calibration and scoring. The chosen item parameters where in line with those typically used and reported in many studies dealing with real and simulated data.

The fact that simulated data generally behave too well should not weaken the value of the findings obtained because the focus here was methodological. However, test developers should expect some decrease in performance between the development phase where simulations are conducted to determine and set up the most appropriate test assembly procedure possible and the operational phase where testing is implemented.

Although the simulated testing situation was typical of many, and the studied test assembly included some of the most commonly used and most promising procedures, the number of CAT algorithms and procedures that are available or in

development is very large and growing. Thus, many other choices could be made that might produce equivalent or better results in a variety of testing situations.

Also, not all factors that may affect the quality of the produced tests were investigated. The overall distributions of item characteristics (more or less discriminating items, for example) and the test length (shorter, larger, and variable test length) are examples of factors that could have an effect on the variability of test information and consequently on the efficiency of the test assembly.

5.2 Directions for Further Research

By explicitly incorporating more and more of the test objectives into the test assembly process, high-stakes on-demand computerized adaptive tests have been greatly improved. However, more remain to be done to improve the ease of implementation, the quality, and the efficiency of the current procedures. Important directions for further research include the development of:

1. more comprehensive and flexible optimization approaches that in addition to test information also take item exposure, testing time, and other important considerations into account in the item selection process, instead of simply increasing the number of constraints imposed onto the system. assembly (Robin, 1999a; Stocking & Swanson, 1993; van der Linden, Scrams & Schnipke, 1999; Veldkamp, 1999); and

2. heuristic (Chang & Ying, 1999; Davey & Fan, 2000; Stocking & Swanson, 1993; Swanson & Stocking, 1993) or mixed integer programming algorithms that will provide optimized solutions to new mathematical programming formulations of the test assembly problem (van der Linden,1998b; van der Linden & Reese,1998; van der Linden & Glas, in press).

5.3 Conclusion

The development of test assembly procedures that will ensure the operational production and delivery of high quality tests on-demand depends on the formulation of clear testing objectives, appropriate test assembly methodologies, and extensive simulations. The main findings of this study are that, with some of the most prominent test assembly procedures:

1. Minimum test information requirements should be specified to ensure that all examinees are provided sufficient opportunity to demonstrate their ability;
2. Conditional item exposure control should be employed to maintain a high level of test security;
3. Item exposure control lowers test measurement to some degree but at the same time greatly improve testing efficiency;
4. A minimum pool size is necessary to ensure the realization of all testing objectives for all examinees; and finally

5. An optimum pool size appears to exist beyond which no gains in testing efficiency are obtained.

On-demand computerized adaptive testing is a very attractive way to conduct testing. It has the potential to offer a very flexible and efficient way to automatically administer tests and report scores. However, this does not come free. Even when using known methodologies, important item and test development efforts have to be made before delivery systems can be used operationally. Once testing is operational, on-going quality control and maintenance work has to be done.

BIBLIOGRAPHY

American Council on Education (1995). Guidelines for computerized-adaptive test development and use in education. Washington, DC: American Council on Education.

Armstrong, R. D., Jones, D. H., & Cordova, M. (1997, March). Mathematical programming approaches to computerized adaptive testing. Paper presented at the annual meeting of the American Educational Research Association, Chicago, Il.

Chang, H. H., & van der Linden, W. J. (2000, April). A zero-one programming model for optimal stratification of item pools in a-stratified computerized adaptive testing. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Chang, H. H., & Ying Z. (1996). A global information approach to CAT. Applied Psychological Measurement, 20, 213-229.

Chang, H. H., & Ying, Z. (1999). A-stratified multi-stage computerized adaptive testing. Applied Psychological Measurement, 23, 211-222.

Chang, S., Ansley, T. N., & Lin, S. H. (2000, April). Performance of item exposure control methods in computerized adaptive testing: Further explorations. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Chang, S., & Twu, B. (1998). A comparative study of item exposure control methods in computerized adaptive testing (Research Rep. No. 98-3). Iowa City, Iowa: American College Testing.

Chen, S., Ankenmann, R. D., & Spray, J. A. (1999, April). Exploring the relationship between exposure rate and test overlap rate in computerized adaptive testing. Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, Québec, Canada.

Cordova, M. J. (1997). Optimization methods in computerized adaptive testing. Unpublished doctoral dissertation, Rutgers University, New Brunswick, NJ.

Davey, T., & Fan, M. (2000, April). Specific information item selection for adaptive testing. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Davey, T., & Nering, M. (1998, September). Controlling item exposure and maintaining item security. Paper presented at the CTB Colloquium: Building the Foundation for Future Assessments. Philadelphia, PA.

Davey, T., & Parshall, C. G. (1995, April). New algorithms for the item selection and exposure control with computerized adaptive testing. Paper presented at the meeting of the American Educational Research Association, San Francisco.

Drasgow, F. & Olsen-Buchanan (Eds) (1999). Innovations in computerized assessment. Mahwah, NJ: Lawrence Erlbaum.

Eignor, D. R., Way, W. D., Stocking, M. L., & Steffen, M. (1993). Case studies in computer adaptive test design through simulation (Research Rep. No. 93-56). Princeton, NJ: Educational Testing Service.

Foster, D. F., Olsen, J., Ford, J. & Sireci, S. G. (1997, March). Administering computerized certification exams in multiple languages: Lessons learned from the marketplace. Paper presented at the meeting of the National Council on Measurement in Education, Chicago.

Gibson, W. M., & Weiner, J. A. (1998). Generating random parallel test forms using CTT in a computer-based environment. Journal of Educational Measurement, 35, 297-310.

Green, B. F. (1983). The promise of tailored tests. In H. W. Wainer and S. Messick (Eds.), Principals of Modern Psychological Measurement, pp. 69-80. Hillsdale, NJ: Lawrence Erlbaum Associates.

Guo, F., Way, W. D. & Reshetar, R. (2000, April). Test security and the development of computerized tests. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Hambleton, R. K. (in press). New computer-based technical issues: Developing items, pretesting, test security, and item exposure. In C. N. Mills, J. Fremer, M. Potenza, W. Ward (Eds.). Computer-based testing: Building the foundation for future assessments. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: principles and applications. Boston, MA: Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Hambleton, R. K., Zaal, J. N., & Pieters J. P. M. (1991). Computerized adaptive testing: Theory, applications, and standards. In R. K., Hambleton, & J. N., Zaal (Eds.). Advances in educational and psychological testing. Boston, MA: Kluwer Academic Publishers.

Hetter, R. D., & Sympson, J., B., (1997). Item exposure control in CAT-ASVAB. In W.A. Sands, B. K. Waters, & J. R. McBride (Eds.). Computerized Adaptive Testing, From Inquiry to Operation (pp. 141-144). Washington, DC: American Psychological Association.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.). Computer-assisted instruction, testing, and Guidance (pp. 139-183). New York, NY: Harper and Row.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. Applied Psychological Measurement, 1, 95-100.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Luecht, R. M., & Nungester R. J. (1998). Some practical examples of computer-adaptive sequential testing. Journal of Educational Measurement, 35, 229-249.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), New horizons in testing (pp. 223-226). New York, Academic Press.

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. Applied Measurement in Education, 9, 287-304.

National Council on Measurement in Education Ad Hoc Committee on Computerized Adaptive Test Item Disclosure (1996). Item disclosure for computerized adaptive tests. Washington, DC: National Council on Measurement in Education.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 70, 351-356.

Parshall C. G. (1998, September). Item development and pretesting in computer-based testing environment. Paper presented at the CTB Colloquium: Building the Foundation for Future Assessments. Philadelphia, PA.

Parshall C. G., Davey, T., & Pashley, P. (in press). Innovative item types for computerized testing. In W. J. van der Linden & C. A. W. Glas (Eds.). Computerized adaptive testing: Theory and practice. Boston, MA: Kluwer Academic Publishers.

Patsula, L. N. (1999). A comparison of computerized adaptive testing and multi-stage testing. Unpublished dissertation, University of Massachusetts, Amherst, MA.

Patsula, L. N., & Steffen, M. (1997, March). Maintaining item and test security in a CAT environment: A simulation study. Paper presented at the meeting of the National Council on Measurement in Education, Chicago.

Revuela, J., & Ponsoda V. (1998). A comparison of item exposure control methods in computerized adaptive testing. Journal of Educational Measurement, 35, 311-327.

Robin, F. (1999a). Alternative item selection strategies for improving test security and pool usage in computerized adaptive testing. Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, Québec, Canada.

Robin, F. (1999b). CBTS: Computer-based testing simulation and analyses [computer program]. Amherst, MA: University of Massachusetts, Laboratory of Psychometric and Evaluative Research.

Robin, F., & Xing, D. (1999). Current and Future Research in Test Security and Item Exposure (Research Report No. 373). Amherst, MA: University of Massachusetts, Laboratory of Psychometric and Evaluative Research.

Sands, W.A., Waters, B. K. & McBride, J. R. (Eds.) (1997). Computerized adaptive testing, from inquiry to operation. Washington, DC: American Psychological Association.

Sireci, S. G., Foster, D. F., Olsen, J., & Robin, F. (1997, March). Comparing dual-language versions of an international certification exam. Paper presented at the meeting of the National Council on Measurement in Education, Chicago.

Stocking, M. L., & Lewis, C. (1995). A new method for controlling item exposure in computerized adaptive testing (Research Rep. No. 95-25). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. Journal of Educational and Behavioral Statistics, 23, 57-75.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. Applied Psychological Measurement, 17, 277-292.

Stocking, M. L., & Swanson, L. (1998). Optimal design of item banks for computerized adaptive tests. Applied Psychological Measurement, 22, 271-279.

Stocking, M. L., Swanson, L. & Perlman, M. (1993). Application of an automated item selection method to real data. Applied Psychological Measurement, 17, 167-176.

Swanson, L. & Stocking, M. L. (1993). A model and heuristic for solving very large selection problems. Applied Psychological Measurement, 17, 151-166.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item exposure rates in computerized adaptive tests. Paper presented at the Annual Conference of the Military Testing Association. San Diego, CA: Military Testing Association.

Theunissen, T. J. J. M. (1985). Binary programming and test design. Psychometrika, 50, 411-420.

van der Linden, W. J. (1998a). Bayesian item selection criteria for adaptive testing. Psychometrika, 63, 201-216.

van der Linden, W. J. (1998b). Optimal assembly of Psychological and educational tests. Applied Psychological Measurement, 22, 195-211.

van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. Psychometrika, 17, 237-247.

van der Linden, W. J., & Glas, C. A. W. (Eds) (in press). Computerized adaptive testing: Theory and practice. Boston, MA: Kluwer Academic Publishers.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. Applied Psychological Measurement, 22, 259-270.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. Applied Psychological Measurement, 23, 195-210.

van der Linden, W. J., Veldkamp, B. P. & Reese, L. M. (2000). An integer programming approach to item pool design. Applied Psychological Measurement, 24, (In press).

Veldkamp, B. P. (1999). Multiple objective test assembly problems. Journal of Educational Measurement, 36, 253-266.

Veldkamp, B. P., & van der Linden, W. J. (1999). Designing item pools for computerized adaptive testing (Research Report 99-03). Enschele, The Netherlands. University of Twente, Faculty of Educational science and Technology.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B., F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). Computerized adaptive testing: A primer. Hillsdale, NJ: Erlbraum.

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. Journal of Educational Measurement, 35, 109-135.

Way, W. D., & Steffen, M. (1998, April). Strategies for managing item pools to maximize item security. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Way, W. D., Steffen, M., & Anderson, D. S. (1998, September). Developing, maintaining, and renewing the item inventory to support computer-based testing. Paper presented at the CTB Colloquium: Building the Foundation for Future Assessments. Philadelphia, PA.

Weiss, D. J. (1973). The stratified adaptive computerized ability test (Research report No. 73-3). Mineapolis, MN: University of Minnesota, Department of Psychology. Psychometric Methods Program.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement, 6, 473-492.

Weiss, D. J. (Ed.). (1983). New horizons in testing: Latent trait test theory and computerized adaptive testing. (pp. 257-283). New York: Academic Press.

Zara, A. R. (1994, March). An overview of the NCLEX/CAT beta test. Paper presented at the meeting of the American Educational Research Association, New Orleans.

Zara, A. R. (1997, March). Administering and scoring the computerized adaptive test. Paper presented at the meeting of the National Council on Measurement in Education, Chicago.

Zenisky, A. (2000). Technological innovations in performance assessment for licensure and certification exams: Current research and future directions (Research Report No. 383). Amherst, MA: University of Massachusetts, Laboratory of Psychometric and Evaluative Research.